REMARKS

The Applicants respectfully request reconsideration of the present application in view of the reasons that follow.

I. Status of the claims

No claims are added, canceled or amended in this paper. Accordingly, claims 5 and 6 are pending and under examination.

II. Claim rejection – non-statutory, obviousness-type double patenting

Claims 5-6 stand rejected on the grounds of non-statutory obviousness-type double patenting over claims 1-9 of U.S. Patent No. 5,851,999 ("the '999 Patent"). The Office Action asserts that claims directed to a *pharmaceutical composition* comprising an expression vector encoding a truncated FLK-1 polypeptide and a pharmaceutically acceptable carrier renders obvious claims directed to *a cell line* comprising a recombinant vector encoding a truncated FLK-1 polypeptide. The Applicants respectfully traverse this ground for rejection.

As noted in the previous reply, the proposed modification would change the principle of operation of the pharmaceutical composition. Additionally, a cell line is not an obvious variant or modification of a pharmaceutical composition.

A. The proposed modification would change the principle of operation of the pharmaceutical composition

Claims 1-9 of the '999 Patent relate to a pharmaceutical composition comprising an expression vector encoding a truncated FLK-1 polypeptide, and a pharmaceutically acceptable carrier. A pharmaceutical composition comprising an expression vector would have a different principle of operation than a pharmaceutical composition including a cell line. For example, a pharmaceutical composition comprising an expression vector would be administered to a patient, (likely targeted to a specific cell population) for uptake by the patient's cells. Once within the

targeted cell population, the encoded truncated polypeptide would be expressed *within* these specific cells. The cells expressing the truncated polypeptide would exhibit the dominant negative phenotype (*e.g.*, inhibit the cellular effects of VEGF binding). In contrast, if a foreign cell line expressing FLK-1 was administered to a patient, the foreign cell line would express the polypeptide outside the target cells. The patient cells would have to somehow "uptake" the truncated polypeptide to illicit the dominant negative affect (*i.e.*, inhibit the cellular effects of VEGF binding).

B. A cell line is not an obvious variant or modification of a pharmaceutical composition

Contrary to the Office Action assertions, one skilled in the art would not find a cell line comprising a recombinant vector to be obvious in light of a pharmaceutical composition comprising an expression vector and a pharmaceutically acceptable carrier. First, a pharmaceutical composition comprising an expression vector and a pharmaceutically acceptable carrier is clearly intended for administration to a patient. A cell line comprising a recombinant vector may not be intended for administration to a patient.

Second, it is possible that administration of a cell line, as opposed to the pharmaceutical composition, could be harmful to the patient and may not yield therapeutic results. For example, it is likely that administering a foreign cell line to a patient would illicit an immune response from the patient. If the patient is suffering from cancer and has undergone chemotherapy, there is a high probability that the patient's immune system is already weakened. Challenging the patient by administering foreign cells would likely cause the patient to become even weaker. Furthermore, a cell line administered to a patient may not even express, or be capable of expressing, the truncated polypeptide after administration. Thus, one of ordinary skill in the art would not consider a cell line including a recombinant vector obvious in light of a pharmaceutical composition comprising an expression vector and a pharmaceutically acceptable carrier.

Accordingly, for at least these reasons, reconsideration and withdrawal of the obviousness-type double patenting rejection is respectfully requested.

III. Claim rejection – 35 U.S.C. § 103(a)

Claims 5-6 are rejected under 35 U.S.C. § 103(a) as allegedly being unpatentable over U.S. Patent No. 5,185,438 ("Lemischka"), Matthews et al., *PNAS*, *88*:9026 (1991) ("Matthews"), and Terman et al., *BBRC*, *187*:1579 (1992) ("Terman"), in view of Ullrich et al., *Cell*, *61*:203 (1990) ("Ullrich"), and Ueno et al. (*Science*, *252*: 844 (1991) ("Ueno-1") and *JBC*, *267*:1470 (1992) ("Ueno-2")). The Office Action assert that because Flk-1 allegedly shares some aspects of homology to PDGFR, FGFR1 and EGFR, that the "invention follows the teaching of the prior art, and achieves exactly the result that would be predicted on the basis of the prior art," and that "there are no unexpected results." Office Action at page 5. Thus, the Office Action alleges that not only would the skilled artisan find it obvious to make the claimed mutant polypeptide (amino acids 1-806 of SEQ ID NO: 2), but that it would not be surprising that the claimed mutant polypeptide would fold into a conformation that could dimerize with a wild-type Flk-1 polypeptide; that the dimer would bind VEGF; and that the dimer would prevent VEGF signaling. The Applicants respectfully traverse this ground for rejection and assert that only with impermissible hindsight would the skilled artisan consider the claimed invention "obvious."

A. Protein function cannot accurately be predicted based on structural or functional homology

Assigning protein function based on sequence homology is viewed with skepticism in the art. For example, Wells illustrates that changes in amino acid sequence, even a change of a few amino acids, can result in proteins with unpredictable function (Wells, *Biochemistry*, *29*(37) 8509-17 (1990), EXHIBIT A). Moreover, Attwood (Attwood, *Science*, *290*: 471-473 (2000), EXHIBIT B) teaches that "[i]t is presumptuous to make functional assignments merely on the basis of some degree of similarity between sequences." Similarly, Skolnick et al. (Skolnick et al., *Trends in Biotech.*, *18*(1): 34-39 (2000), EXHIBIT C) teach that the skilled artisan is well

aware that assigning functional activities for any particular protein or protein family based on sequence homology is inaccurate, in part because of the multifunctional nature of proteins (*see e.g.*, Skolnick et al. at Abstract and "sequence-based approaches to functional prediction," page 34).[1]

In the present situation, there is very little homology between these polypeptides at the amino acid level. Exhibits D-F show protein BLAST alignments of SEQ ID NO: 2 amino acids 1-806 against PDGFR (human; Accession Number NP_002600, EXHIBIT D), EGFR (human; Accession Number AAH94761, EXHIBIT E), and FGFR1 (human; Accession Number AAH15035, EXHIBIT F) and as identified on the NCBI database. The alignments make clear that sequence homology between these polypeptides is minimal. Accordingly, the skilled artisan would not assume or reasonably believe that the truncated form of Flk-1 would behave in a manner similar to a truncated form of PDGFR, EGFR or FGFR1.

Even in situations where there is some confidence of similar overall structure between two proteins, only experimental research can confirm the artisan's best guess as to the function of the structurally related proteins (*see e.g.*, Skolnick, in particular Abstract and Box 2). It is well known in the art that predicting protein function from sequence data is extremely complex. While it may be that one or more deletions are generally possible in any given protein, the positions within the protein's sequence where such deletions can be made with a reasonable expectation of producing a desired or expected function should be determined empirically for each protein. Certain positions within a sequence are critical to the protein's structure/function relationship, *e.g.*, such as various sties or regions directly involved in binding, activity and in providing the correct three-dimensional spatial orientation of binding and active sites (*see e.g.*, Wells).

[1] MPEP § 2124 states that "in some circumstances a factual reference need not antedate the filing date," for example, when describing a "scientific truism." Both Attwood and Skolnick present such scientific truisms.

Accordingly, the Applicants respectfully assert that even in light of the combination of cited references, the skilled artisan would not have expected the claimed polypeptide – a deletion mutant of Flk-1 with very little amino acid homology to EGFR, FGFR or PDGFR – to form a three-dimensional structure capable of dimerizing with a wild-type Flk-1 polypeptide, function with the wild-type polypeptide to bind VEGF, and to inhibit VEGF signaling function. As such, there would be no reason for the skilled and creative artisan to try to generate the claimed Flk-1 mutant, and there would be no reasonable expectation that such a mutant would even be functional, yet alone be beneficial. The Applicants respectfully assert that the claimed cell lines are not obvious in light of the combination of cited references.

B. None of the cited references disclose that Flk-1 functions as a dimer; accordingly, activity of a Flk-1 mutant can not be predicted based on the activity of kinase receptor dimer mutants

The Office asserts that Terman teaches "that it would be desirable to investigate the dimeric combinations in which the receptor occurs; Terman does not doubt that the receptor is a dimer." Office Action at page 4. The Applicants respectfully disagree with the Examiner's characterizations regarding Terman. In fact, Terman states that *"[i]t is not known* whether KDR and *flt* can form functionally active dimers analogous to the PDGF receptor dimers" and "*it is not known* whether KDR, *flt* or heterodimer KDR/*flt* mediates mitogenic activity and/or vascular permeability." Terman at page 1585. Accordingly, the teaching of Terman does not substantiate the Examiner's assertions that Flk-1 was known to function as a dimer.

The Office also asserts that "[w]ith respect to Ullrich, combination of the subunit of SEQ ID NO: 2 *in vivo* in a cell that expresses Flk-1 would inherently result in a combination with a "normal" subunit, regardless of whether the receptor were a homo- or hetero-dimer." Office Action at page 4. However, the claimed SEQ ID NO: 2 mutant is neither taught nor suggested in Ullrich. Accordingly, this argument is moot with respect to the claimed invention.

The Office Action also asserts that Ullrich teaches that "[r]eceptor oligomerization is a universal phenomenon among growth factor receptors." Office Action at page 4. Ullrich does indeed make this statement. However, this statement, alone or in combination with any or all of the cited references, does not in any way teach or suggest that the claimed mutant polypeptide would be able to "oliogmerize."

The Office Action also cites Ullrich with respect to the function of various kinase mutants: "While the kinase activity of the various receptors was dispensable for their expression and targeting to the cell surface it was indispensable for signal transduction and induction of other early and delayed cellular responses...." Office Action at page 4. Based on this statement, the Office Action concludes "without knowing the subunit structure of the receptor, one would expect that a subunit lacking the tyrosine kinase domain would have dominant negative signaling effects." Office Action at page 5. The Applicants respectfully disagree with the characterization of Ullrich and the conclusion reached. While Ullrich describes mutations of the EGF, insulin and PDF receptors, Ullrich does not teach or suggest Flk-1 or Flk-1 mutations. Moreover, Ullrich in combination with the cited references does not teach or suggest that wild-type Flk-1 functions as a dimer; accordingly, there is no reason one of skill in the art would conclude that the claimed mutant Flk-1 would form a dimer with its wild-type counterpart. Further, one of ordinary skill in the art would not conclude that "without knowing the subunit structure of the receptor, one would expect that a subunit [of Flk-1] lacking the tyrosine kinase domain would have dominant negative signaling effects."

With respect to mutant dimerization and mutant-dimer function, the Office Action asserts that "Applicant are arguing limitations that are not found in the claims," and that "neither a dimer, nor any particular activity" is required in the claims. Office Action at page 5. The Applicants respectfully traverse this assertion.

The Applicants assert that because the ability of wild-type Flk-1 to dimerize was *unknown* at the time of filing, the fact that the claimed mutant Flk-1 could dimerize, that a dimer including the claimed mutant Flk-1 would bind VEGF, and that the a dimer including the claimed mutant Flk-1 would inhibit VEGF signaling *is completely unexpected*.

For at least these reasons, the Applicants respectfully disagree with the Office Action assertions that one skilled in the art would find it obvious to make the truncated mutants, and to expect the claimed truncated polypeptide "to render endogenous wild-type Flk-1 unresponsive to VEGF and inhibit the cellular effects of VEGF binding." The Applicants respectfully contend that, for the claimed Flk-1 sequence, dominant negative inhibition of cellular effects of VEGF binding *would not have been obvious* to a skilled artisan. Specifically, one of ordinary skill in the art would not have expected the claimed polypeptide to "render endogenous wild-type Flk-1 unresponsive to VEGF" as recited in the claims *because even the wild-type Flk-1 was not know to function as a dimer*.

Accordingly, reconsideration and withdrawal of the rejection under 35 U.S.C. § 103(a) is respectfully requested.

C. Because *flt* was a known VEGF receptor with high affinity for VEGF, it was unexpected that the *flk-1* mutant would inhibit the cellular effects of VEGF binding

In addition, it was entirely unexpected that the truncated Flk-1 variant would have an inhibitory effect on the cellular response of VEGF. Such results were unexpected because at least one other receptor, *flt*, was known to bind VEGF with high affinity.[2] It also was known that *flt* is expressed in endothelial cells of a growing tumor.[3] Significantly, *flt* has a 50-fold higher affinity for VEGF than Flk-1.[4] Given the importance of VEGF signaling in angiogenesis during *inter alia*,

[2] *See e.g.,* Terman et al., *BBRC*, *187*:1579 (1992).

[3] *See*, Plate et al., *Nature*, *359*: 845-848 (1992); Plate et al., *Cancer Research*, *53*: 5822-5827 (1993). (Ref. A35)

[4] *See*, Waltenberger et al., *J. Biol. Chem.*, *269*: 26988-26995 (1994).

development, wound healing and organ regeneration, some redundancy in the system would be expected. Consequently, the skilled artisan would not have expected that blocking the Flk-1 signaling pathway would shut down the cellular response to VEGF, resulting in suppression of angiogenesis and inhibition of tumor growth. Rather, one of ordinary skill in the art would have anticipated that the biological response to VEGF, such as the proliferation of blood vessels, would still be transduced through *flt* or some as yet undiscovered receptors.

For at least these reason, the ability of the claimed truncated Flk-1 receptor proteins to inhibit angiogenesis were unexpected. Accordingly, reconsideration and withdrawal of the rejection is respectfully requested

IV. Conclusion

The Examiner is invited to contact the undersigned by telephone if it is felt that a telephone interview would advance the prosecution of the present application.

The Commissioner is hereby authorized to charge any additional fees which may be required regarding this application under 37 C.F.R. §§ 1.16-1.17, or credit any overpayment, to Deposit Account No. 19-0741. Should no proper payment be enclosed herewith, as by the credit card payment instructions in EFS-Web being incorrect or absent, resulting in a rejected or incorrect credit card transaction, the Commissioner is authorized to charge the unpaid amount to Deposit Account No. 19-0741.

If any extensions of time are needed for timely acceptance of papers submitted herewith, the Applicants hereby petition for such extension under 37 C.F.R. § 1.136 and authorize payment of any such extensions fees to Deposit Account No. 19-0741.

Respectfully submitted,

Date: February 18, 2010

By: /Stephanie H. Vavra/ Reg. No. 45,178

FOLEY & LARDNER LLP
Customer Number: 22428
Telephone: (202) 672-5538
Facsimile: (202) 672-5399

for Michele M. Simkin
Attorney for the Applicants
Registration No. 34,717

Biochemistry

 Volume 29, Number 37 September 18, 1990

Perspectives in Biochemistry

Additivity of Mutational Effects in Proteins

James A. Wells

Protein Engineering Department, Genentech, Inc., 460 Point San Bruno Boulevard, South San Francisco, California 94080

Received April 19, 1990; Revised Manuscript Received May 29, 1990

The energetics of virtually all binding functions in proteins is the culmination of a set of molecular interactions. For example, removal of a single molecular contact by a point mutation causes relatively small reductions (typically 0.5–5 kcal/mol) in the free energy of transition-state stabilization [for reviews see Fersht (1987) and Wells and Estell (1988)], protein–protein interactions (Laskowski et al., 1983, 1989; Ackers & Smith, 1985), or protein stability [for review see Matthews (1987)] compared to the overall free energy associated with these functional properties (usually 5–20 kcal/mol). Thus, it is possible to modulate protein function by mutation at many contact sites. In fact, to design large changes in function will often require mutation of more than one functional residue.

There is now a large data base for free energy changes that result when single mutants are combined. A review of these data shows that, in the majority of cases, the sum of the free energy changes derived from the single mutations is nearly equal to the free energy change measured in the multiple mutant. However, there are two major exceptions where such simple additivity breaks down. The first is where the mutated residues interact with each other, by direct contact or indirectly through electrostatic interactions or structural perturbations, so that they no longer behave independently. The second is where the mutation causes a change in mechanism or rate-limiting step of the reaction. It is important to note that the additive effects discussed here do not change the molecularity of their respective reactions. When the molecularity of the reaction changes [as in comparing the free energy of binding of one linked substrate (A–B) versus the sum of two fragments (A plus B)], large deviations from simple additivity can result from entropic effects (Jencks, 1981). Although the focus here is on enzyme activity, similar conclusions may be drawn from mutations affecting protein–protein interactions, protein–DNA recognition, or protein stability. Some practical examples and applications are discussed.

Additivity Relationships

The change in free energy of a functional property caused by a mutation at site X is typically expressed relative to that of the wild-type protein as $\Delta\Delta G_{00}$. Such free energy changes for two single mutants (X and Y) can be related to those of a double mutant (designated X,Y) by eq 1 (Carter et al., 1984; Ackers & Smith, 1985). The ΔG_i term (also called the coupling energy; Carter et al., 1984) should reflect the extent to which the change in energy of interaction between sites X and Y affects the functional property measured. It is possible for ΔG_i to be either positive or negative depending upon whether the interactions between the mutant side chains reduce or enhance the functional property measured. Furthermore, the ΔG_i term should not exceed the free energy of interaction between side chains at sites X and Y except in cases where these mutations cause large structural perturbations. This was first applied to evaluating the functional independence of residues mutated in tyrosyl-tRNA synthetase (Carter et al., 1984). In one case the sum of the $\Delta\Delta G$ values for single mutants was equal to that of the double mutant, indicating the sites functioned independently; in another example there was a large discrepancy, suggesting the sites were interacting.

$$\Delta\Delta G_{(X,Y)} = \Delta\Delta G_{00} + \Delta\Delta G_{(Y)} + \Delta G_i \qquad (1)$$

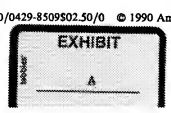
Simple Additivity in Transition-State Binding Interactions

The strengths of noncovalent interactions are strongly dependent upon the nature of the two groups and the distance (r) between them. For example, the free energy of charge–charge, random charge–dipole, random dipole–dipole, van der Waals attraction, and repulsion decay as $1/r$, $1/r^4$, $1/r^6$, $1/r^6$, and $1/r^{12}$, respectively [for review see Fersht (1985)]. Thus, when the side chains at sites X and Y are remote to one another and assuming no large structural perturbations, the ΔG_i term should be negligible and eq 1 thus simplifies to

$$\Delta\Delta G_{(X,Y)} \simeq \Delta\Delta G_{00} + \Delta\Delta G_{(Y)} \qquad (2)$$

This situation, here referred to as simple additivity, is generally observed except where side chains are close to each other or when one or both of the mutants change the rate-limiting step or reaction mechanism. These principles are well illustrated from data of additive mutational effects on transition-state stabilization energies.

0006-2960/90/0429-8509$02.50/0



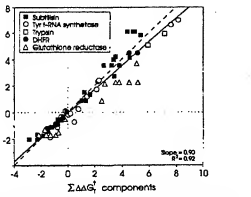


FIGURE 1: Plot of the changes in transition-state stabilization energies for the multiple mutant versus the sum for the component mutants. Data are taken from Table I and represent mutants from subtilisin (■), tyrosyl-tRNA synthetase (○), trypsin (□), DHFR (●), and glutathione reductase (△), where mutant or wild-type side chains should not contact one another. The dashed line has a slope of 1, and the solid line is a best fit to all the data.

Changes in transition-state stabilization energy ($\Delta\Delta G_T^{\ddagger}$) caused by a mutation can be calculated from eq 3 (Wilkinson et al., 1983), in which R is the gas constant, T is the absolute

$$\Delta\Delta G_T^{\ddagger} = -RT \ln \frac{(k_{cat}/K_M)_{mutant}}{(k_{cat}/K_M)_{wild\text{-}type}} \qquad (3)$$

temperature, k_{cat} is the turnover number, and K_M is the Michaelis constant for the mutant and wild-type enzyme against a fixed substrate. $\Delta\Delta G_T^{\ddagger}$ represents the change in free energy to reach the transition-state complex (E·S*) from the free enzyme and substrate (E + S).

To analyze the proposition that the interaction energy term, $\Delta G_{T(I)}^{\ddagger}$, is relatively small when the sites of mutation (X and Y) are remote to one another, $\Delta\Delta G_T^{\ddagger}$ values were collected from the literature where side-chain substitutions in the multiple mutant are beyond van der Waals contact (>4 Å distant) from each other (Table I). There are at least 25 examples distributed across five different enzymes where $\Delta\Delta G_T^{\ddagger}$ values can be calculated for the individual and multiple mutants assayed in at least two different ways. Among these are examples where electrostatic interactions, hydrogen bonding, and steric and hydrophobic effects have been altered separately or in combination with others. The X-ray structures of the wild-type proteins show that the wild-type side chains are not in contact. Modeling suggests the mutant side chains are beyond possible van der Waals contact unless the mutant side chains were to cause significant changes in the overall protein structure. Such large changes are rarely observed in structures of site-specific mutant proteins (Katz & Kossiakoff, 1986; Alber et al., 1987; Howell et al., 1986; Wilde et al., 1988) or even highly variant natural proteins (Chothia & Lesk, 1986).

A collective plot of the sum of the $\Delta\Delta G_T^{\ddagger}$ values for the component mutants versus the corresponding multiple mutant (Table I) gives a remarkably strong correlation ($R^2 = 0.92$) with a slope near unity (Figure 1). The simplest interpretation is that the interaction term, $\Delta G_{T(I)}^{\ddagger}$, is small compared to the overall effects on $\Delta\Delta G_{T(X,Y)}^{\ddagger}$. It is formally possible that there are large and compensating effects between side chains X and Y that systematically lead to small net values for $\Delta G_{T(I)}^{\ddagger}$.

There are some notable exceptions that weaken the correlation within the data set (Table I). In particular, combining the R204L mutation in *Escherichia coli* glutathione reductase gives a less than additive effect, especially when combined with another mutant, R198M (Scrutton et al., 1990). These basic residues are not in direct contact, but both side chains form a salt bridge with the 2′-phosphate group of NADPH. Indeed, the largest discrepancies are when these mutants are assayed with NADPH as compared to NADH. Similarly, the sum of the $\Delta\Delta G_T^{\ddagger}$ values for two positively charged component mutants in subtilisin (D99K and E156K) overestimates the effect of the multiple mutant when assayed with an Arg but not with a Phe substrate (Russell & Fersht, 1987). Such discrepancies are not too surprising because charge–charge interactions fall off as $1/r$ and can exhibit long-range effects in proteins [for example, see Russell and Fersht (1988)]. The physical basis for other large discrepancies not involving electrostatic substitutions is less clear but may involve unexpectedly large structural changes or changes in enzyme mechanism (see below).

These additivity tests are not particularly dominated by one of the single mutants in the sum. The average contribution (±SE) for the most dominant mutant in each sum calculated from the 69 additivity tests given in Table I is only 68% (±15%) of the total sum (theoretical is ~50%). Furthermore, the plot in Figure 1 is not analogous to graphs of correlated variables, where A is plotted versus the sum of A + B, because in Figure 1 the values on the y-axis are determined *independently* from those on the x-axis.

Complex Additivity in Transition-State Stabilization—When $\Delta G_{T(I)}^{\ddagger} \neq 0$

(A) Change in Interaction Energy between Sites X and Y. Where residues X and Y are close enough to contact, it is more likely that the $\Delta G_{T(I)}^{\ddagger}$ term will be significant. There are 11 examples collectively from tyrosyl-tRNA synthetase and subtilisin that fit this category (Table II).

A series of mutants in tyrosyl-tRNA synthetase at positions 48 and 51 (Carter et al., 1984; Lowe et al., 1985) show complex additivity (Table II). His48 and Thr51 in the wild-type structure are next to each other on adjacent turns of an α-helix. His48 hydrogen bonds to the ribose ring oxygen of ATP while Thr51 can make van der Waals contact with ATP. The T51P mutation increases the catalytic efficiency of the enzyme in some assays by more than −2 kcal/mol (Wilkinson et al., 1984). However, when this mutation is combined with mutations at position 48, the effects are not simply additive. An X-ray structure of the T51P mutant indicates there are no structural changes in the α-helix (Brown et al., 1987). Instead, it is suggested that the T51P mutant is improved over wild type because the wild-type enzyme contains a bound water in the vicinity of Thr51 that disfavors substrate binding. Blow and co-workers (Brown et al., 1987) argue that the change in solvent structure propagated to position 48 may account for the complex additivity. In the previous section, the double mutant (H48G,T51A) exhibited nearly simple additivity (Table I). Presumably, the smaller and less hydrophobic alanine substitution at position 51 should not introduce as large a change in solvent structure as the pyrrolidone ring of proline.

In the case of subtilisin (Table II), Glu156 is near the top of the P1 binding crevice while Gly166 is at the bottom. In the wild-type enzyme these sites do not make direct van der Waals contact, but large side chains substituted at position 166 can be modeled to contact the residue at position 156. In fact, X-ray structural analysis shows that an Asn side chain at position 166 makes a good hydrogen bond with Glu156 (Bott et al., 1987). Moreover, all of the substitutions are polar or charged, the energetics of which are expected to be the most long range. Thus, the mutant side chains alter substantially the intramolecular interactions between positions 156 and 166.

Table I: Comparison of Sums of $\Delta\Delta G_T$[a] from Component Mutants vs the Multiple Mutant Where the Mutant or Wild-Type Side Chains *Do Not* Contact One Another

assay	$\Delta\Delta G_T$[b] component mutants		sum	multiple mutant
	Tyrosyl-tRNA Synthetase			
	C35G + H48G[d]			
ATP/PP_i	+1.20	+1.04	+2.24	+2.30
ATP/tRNA	+1.05	+1.13	+2.18	+1.68
Tyr/PP_i	+1.14	+1.12	+2.26	+2.32
Tyr/tRNA	+0.32	+1.12	+1.45	+1.20
	C35G + T51P			
ATP/PP_i	+1.20	−1.91	−0.71	−1.14
ATP/tRNA	+1.05	−2.35	−1.30	−1.88
Tyr/PP_i	+1.14	−0.64	+0.50	−0.74
Tyr/tRNA	+0.32	+0.50	+0.82	+0.21
	C35G + T51C[e]			
ATP/tRNA	+1.05	−0.93	+0.12	−0.22
ATP/Tyr	+1.14	−0.91	+0.23	−0.13
	H48N + T51A[e]			
ATP/PP_i	+0.26	−0.38	−0.12	+0.04
ATP/tRNA	−0.13	−0.32	−0.45	−0.37
	T40A + H45G[d]			
Tyr/Tyr	+5.02	+3.15	+8.17	+6.95
ATP/Tyr	+5.13	+2.44	+7.57	+6.67
	Rat Trypsin			
	G216A + G226A[d]			
K	+2.75	+3.13	+5.88	+5.07
R	+2.19	+4.91	+7.10	+5.90
	Dihydrofolate Reductase ($\Delta\Delta G_{binding}$)			
	F31V + L54G[f]			
H_2F	+1.6	+2.9	+4.5	+4.5
MTX	+2.2	+2.9	+5.1	+4.5
	Subtilisin BPN'			
	E156S + Y217L + G169A[g]			
E	−1.43	−0.87	−0.62	−2.92	−2.06
Q	−0.60	−0.36	−0.32	−1.28	−1.14
A	−0.15	−0.41	−0.27	−0.83	−0.92
K	+1.70	−0.08	−0.30	+1.32	+0.87
M	−0.86	−0.32	−0.39	−1.57	−1.41
F	−0.61	−0.29	−0.66	−1.56	−1.17
Y	−0.24	−0.12	−0.41	−0.77	−0.59
	E156S + Y217L				
E	−1.43	−0.87		−2.30	−1.67
Q	−0.60	−0.36		−0.96	−0.96
A	−0.15	−0.41		−0.56	−0.53
K	+1.70	−0.08		+1.62	+1.33
M	−0.86	−0.32		−1.18	−1.11
F	−0.61	−0.29		−0.90	−0.84
Y	−0.24	−0.12		−0.36	−0.32
	E156S, Y217L + G169A				
E	−1.67	−0.62		−2.29	−2.06
Q	−0.96	−0.32		−1.28	−1.14
A	−0.53	−0.27		−0.80	−0.92
K	+1.33	−0.30		+1.03	+0.87
M	−1.11	−0.39		−1.50	−1.41
F	−0.84	−0.66		−1.50	−1.17
Y	−0.32	−0.41		−0.73	−0.59
	D99S + E156S[d]				
R	+0.47	+0.77		+1.24	+1.52
F	0	−0.62		−0.62	−0.52

assay	$\Delta\Delta G_T$[b] component mutants		sum	multiple mutant
	Subtilisin BPN'			
	D99K + E156K			
R	+1.29	+2.12	+3.41	+2.74
F	+0.13	−0.49	−0.36	−0.42
	E156S, G166A + G169A, Y217L[i]			
F	−0.40	−1.46	−1.86	−1.76
Y	+0.94	−1.03	−0.09	+0.02
	G166A + S24C, H64A			
F	−0.40	+4.96	+4.56	+4.11
Y	+0.94	+4.40	+5.34	+5.84
	E156S, G169A, Y217L + S24C, H64A			
F	−1.46	+4.96	+3.50	+4.21
Y	−1.03	+4.40	+3.37	+3.96
	S24C, H64A, G169A, Y217L + G166A			
F	+4.21	−0.40	+3.81	+3.53
Y	+3.96	+0.94	+4.90	+6.07
	S24C, H64A, G166A + E156S, G169A, Y217L			
F	+4.11	−1.46	+2.65	+3.53
Y	+5.84	−1.03	+4.81	+6.07
	S24C, H64A + E156S, G166A, G169A, Y217L			
F	+4.96	−1.76	+3.20	+3.53
Y	+4.40	+0.02	+4.38	+6.07
	***E. coli* Glutathione Reductase**			
	A179G + R198M[j]			
NADH	−1.10	−0.62	−1.72	−1.32
NADPH	+0.08	+2.68	+2.76	+2.11
	A179G + R204L			
NADH	−1.10	+0.41	−0.69	−1.54
NADPH	+0.08	+2.42	+2.50	+0.87
	R198M + R204L			
NADH	−0.62	+0.41	−0.21	−0.51
NADPH	+2.68	+2.42	+5.10	+3.70
	A179G + R179M, R204L			
NADH	−1.10	−0.51	−1.61	−1.72
NADPH	+0.08	+3.70	+3.78	+2.22
	R198M + A179G, R204L			
NADH	−0.62	−1.54	−2.16	−1.72
NADPH	+2.68	+0.87	+3.55	+2.22
	R204L + A179G, R198M			
NADH	+0.41	−1.32	−0.91	−1.72
NADPH	+2.42	+2.11	+4.53	+2.22
	R179G + R198M + R204L			
NADH	−1.10	−0.62	+0.41	−1.31	−1.72
NADPH	+0.08	+2.68	+2.42	+5.18	+2.22

[a] Carter et al. (1984). The assays refer to measurements of ATP-dependent pyrophosphate exchange (ATP/PP_i) or tRNA charging (ATP/tRNA) under saturating conditions for tyrosine and vice versa for Tyr/PP_i exchange and Tyr/tRNA charging. [b] Lowe et al. (1985). The ATP/Tyr activation assay refers to formation of tyrosyl adenylate under saturating concentrations of tyrosine. [c] Jones et al. (1986). [d] Leatherbarrow et al. (1986). The ATP/Tyr and Tyr/Tyr activation assays refer to formation of tyrosyl adenylate under pre-steady-state conditions, and k_{ss}/K_d is calculated from k_1/K_d for tyrosine and ATP, respectively. [e] Craik et al. (1985). The substrate was D-Val-Leu-(X)-aminofluorocoumarin where the P1 residue (X) is either Lys (K) or Arg (R). [f] Mayer et al. (1986). The ligand was either dihydrofolate (H_2F) or methotrexate (MTX). [g] Wells et al. (1987a). The substrate was succinyl-L-Ala-L-Ala-L-Pro-L-(X)-p-nitroanilide where the P1 (X) residue (Sobeckter & Berger, 1937) was either Glu (E), Gln (Q), Ala (A), Lys (K), Met (M), Phe (F), or Tyr (Y). [h] Russell and Fersht (1987). The substrate was benzoyl-L-Val-Gly-L-Arg-p-nitroanilide (R) or succinyl-L-Ala-L-Ala-L-Pro-L-Phe-p-nitroanilide (F). [i] Carter et al. (1989). The substrate was succinyl-L-Phe-L-Ala-L-His-L-(X)-p-nitroanilide where X was either Phe (F) or Tyr (Y). [j] Scrutton et al. (1990). The assay followed the reduction of oxidized glutathione by NADH or NADPH.

Table II: Comparison of Sums of $\Delta\Delta G_t^{\ddagger}$ from Component Mutants vs the Multiple Mutant Where the Mutant Side Chains *Can* Contact One Another

	$\Delta\Delta G_t^{\ddagger}$[b]			
assay[a]	component mutants		sum	multiple mutant
Tyrosyl-tRNA Synthetase				
	H48G + T51P[b]			
ATP/PP_i	+1.04	−1.91	−0.87	+1.07
ATP/tRNA	+1.13	−2.35	−1.22	+0.77
Tyr/PP_i	+1.12	−0.64	+0.48	+1.02
Tyr/tRNA	+1.12	+0.50	+1.63	+0.17
ATP/Tyr[c]	+0.95	−1.99	−1.04	+1.04
Tyr/ATP	+1.07	−0.38	+0.69	+0.82
	H48N + T51P			
ATP/Tyr	+0.18	−1.99	−1.81	−0.76
Tyr/Tyr	+0.36	−0.38	−0.02	−0.64
ATP/tRNA	−0.02	−2.23	−2.25	−1.07
	N48G + T51P			
ATP/Tyr	+0.37	−0.94	−0.57	+0.86
Tyr/Tyr	+0.41	−1.00	−0.59	+0.45
ATP/tRNA	+1.26	−1.05	+0.21	+0.90
	Q48G + T51P			
ATP/Tyr	+1.31	−1.09	−2.40	−1.22
Tyr/Tyr	−2.05	−1.65	−3.70	−2.31
ATP/tRNA	−1.87	−1.85	−3.72	−2.23
	H48Q + T51P			
ATP/Tyr	+2.26	−1.99	+0.27	+1.17
Tyr/Tyr	+3.13	−0.38	+2.75	+1.48
ATP/tRNA	+3.11	−2.23	+0.88	+1.26
Subtilisin BPN'				
	E156Q + G166D[d]			
Q	−1.04	+1.27	+0.23	+0.75
M	−0.45	+1.83	+1.38	+0.16
K	+2.15	+0.53	+2.68	+0.26
	E156S + G166D			
Q	−0.59	+1.27	+0.68	+0.74
M	−0.85	+1.83	+0.98	+0.66
K	+1.68	+0.53	+2.22	+0.49
	E156Q + G166N			
E	−1.71	−0.11	−1.82	−0.69
Q	−1.04	+0.14	−0.90	−0.77
M	−0.45	+0.18	−0.27	−1.10
K	+2.15	+0.48	+2.73	+1.16
	E156S + G166N			
E	−1.44	−0.11	−1.55	−0.51
Q	−0.59	+0.14	−0.45	−0.85
M	−0.85	+0.18	−0.67	−0.78
K	+1.68	+0.48	+2.16	+1.26
	E156S + G166K			
E	−1.44	−3.49	−4.93	−4.49
Q	−0.59	−1.03	−1.62	−0.95
M	−0.85	−1.37	−2.22	−1.12
K	+1.68	+0.51	+2.19	+1.88
	E156Q + G166K			
E	−1.71	−3.49	−5.20	−4.49
Q	−1.04	−1.03	−2.07	−0.95
M	−0.45	−1.37	−1.82	−1.12
K	+2.15	+0.51	+2.66	+1.88

[a] See Table I for description assays. [b] Lowe et al. (1985). [c] Carter et al. (1984). [d] Wells et al. (1987b).

In these six examples there are large and systematic discrepancies between the sum of the $\Delta\Delta G_t^{\ddagger}$ values for the single mutants and those of the corresponding double mutant (Wells et al., 1987b). In almost all cases, the sum of the $\Delta\Delta G_t^{\ddagger}$ values for the single mutants is much greater than the value for the multiple mutant. Nonetheless, the $\Delta\Delta G_t^{\ddagger}$ value predicted from the sum of the single mutants does have the same sign as that for the double mutant, so that the single mutants predict qualitatively the effect on the multiple mutant.

A plot (Figure 2) of the collective data set from Table II is in contrast to that seen in Figure 1. The $\Delta\Delta G_t^{\ddagger}$ values for the multiple mutants correlate more poorly with the sum of

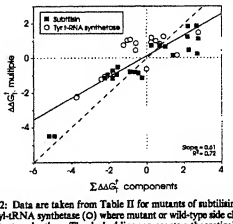


FIGURE 2: Data are taken from Table II for mutants of subtilisin (■) or tyrosyl-tRNA synthetase (○) where mutant or wild-type side chains can contact each other. The dashed line represents a theoretical line of unity slope, and the solid line represents the best fit.

the component single mutants ($R^2 = 0.72$). Moreover, the slope of the line (0.61) is much below unity. This indicates that the function of one residue is compromised by mutation of another. Of the 40 additivity examples, the average contribution of the most dominant single mutant to the sum of the $\Delta\Delta G_t^{\ddagger}$ values is 71% (±13%) of the total. Thus (as in Figure 1), both single mutants can contribute substantially to free energy changes measured in the multiple mutant. However, this data set is derived from mutations at only two different sites on two different proteins.

In summary, complex additivity can be observed when mutations at sites X and Y change the intramolecular interaction energy between sites. This can be mediated by direct steric, electrostatic, hydrogen-bonding, or hydrophobic interactions or indirectly through large structural changes in the protein, solvent shell, or electrostatic interactions. Complex additivity is most likely to occur where the sites of mutation are very close together and larger or chemically divergent side chains are introduced.

(B) Mutations at Sites X or Y Change the Enzyme Mechanism or Rate-Limiting Step. If the catalytic functions of two or more residues are interdependent, then a mutation of one residue can affect the functioning of the other(s). This form of complex additivity is well illustrated for mutations in the catalytic triad and oxyanion binding site of subtilisin (Carter & Wells, 1988, 1990). In the catalytic mechanism of subtilisin (Figure 3), the rate-limiting step in amide bond hydrolysis is transfer of the proton from Ser221 to His64 with nucleophilic attack upon the scissile carbonyl carbon. This is accompanied by electrostatic stabilization of the protonated imidazole by Asp32 and hydrogen bonding to the oxyanion by the side chain of Asn155 and the main-chain amide of Ser221. Mutational analysis shows that once the catalytic Ser221 is mutated to Ala (S221A), additional mutations in the triad or oxyanion binding site cause no further loss in catalytic efficiency (Table III).

The S221A enzyme retains a catalytic activity that is still 10^4 above the solution hydrolysis rate (Carter & Wells, 1988). It is proposed that this residual activity is derived from remaining transition-state binding contacts outside of the catalytic triad coupled with solvent attack upon the carbonyl carbon from the face opposite position 221 (Carter & Wells, 1990). This proposal is based on a model showing that there is no room for a water molecule near Ala221 once the substrate is bound. Furthermore, conversion of Asn155 to Gly enhances the activity of the S221A mutant by −1.2 kcal/mol (Table III).

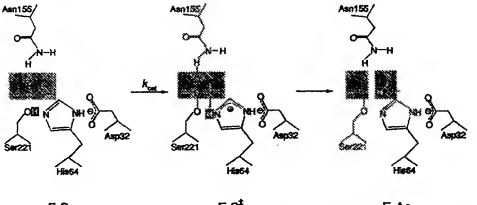


FIGURE 3: Schematic diagram of the mechanism of subtilisin showing the rate-limiting acylation step for hydrolysis of peptide bonds. Reproduced with permission from Carter and Wells (1988). Copyright 1988 Macmillan.

Table III: Comparison of Sums of $\Delta\Delta G_T$[a] from Component Mutants vs the $\Delta\Delta G_T$[a] for Multiple Mutants in the Catalytic Triad and Oxyanion Binding Site of Subtilisin BPN'[a]

component mutants			sum	multiple mutant
S221A + H64A[b]				
+8.93	+8.84		+17.76	+8.83
S221A + D32A				
+8.93	+6.52		+15.45	+8.86
H64A + D32A				
+8.84	+6.52		+15.36	+7.48
S221A + H64A + D32A				
+8.93	+8.84	+6.52	+24.29	+8.65
S221A + H64A, D32A				
+8.93	+7.48		+16.40	+8.65
H64A + S221A, D32A				
+8.84	+8.86		+17.70	+8.65
D32A + S221A, H64A				
+6.52	+8.83		+15.35	+8.65
S221A + N155G[c]				
+8.93	+3.08		+12.01	+7.70

[a] All enzymes were assayed with the substrate succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide. [b] Carter and Wells (1988). [c] Carter and Wells (1990).

This is consistent with the opposite-face solvent attack mechanism of S221A, because the oxyanion (Figure 3) would develop away from Asn155 and the N155G mutation improves solvent accessibility to the scissile carbonyl carbon.

Complex additivity is also seen for subtilisin mutated at positions 64 and 32. The double (H64A,D32A) and corresponding single mutants show a linear dependence upon hydroxide ion concentration (between pH 8 and 10) that may reflect hydroxide assistance in the deprotonation of the Oγ of Ser221 (Carter & Wells, 1988). Thus, once His64 is converted to Ala, Asp32 is a liability, presumably by electrostatic repulsion of hydroxide ion. [Note the −1.3 kcal/mol improvement in $\Delta\Delta G_T^{\ddagger}$ for the double mutant (H64A,D32A) compared to H64A alone; Table III.]

In summary, if an enzyme mechanism relies upon cooperative interaction between two or more residues, then multiple mutations within this subset can result in large values for $\Delta G_{T(1)}^{\ddagger}$. In fact, if the mechanism is changed substantially, residues that were a catalytic asset can become a liability. Simple additivity can also break down when one or more of the mutations cause a change in the rate-limiting step. In an extreme case, one may have a number of mutants in an enzyme that enhance the activity, but the cumulative enhancement of activity could not go beyond the diffusion-controlled limit (Albery & Knowles, 1976).

Additive Effects on Substrate Binding

The analysis above considered changes in binding free energies between the free enzyme and substrate (E + S) to yield the bound transition-state complex (E-S‡). The steady-state kinetic analysis for subtilisin and tyrosyl-tRNA synthetase is such that the K_M values approximate the enzyme–substrate dissociation constant K_s. Additivity analysis based on calculations of $\Delta\Delta G_{binding}$ (from K_M values) or $\Delta\Delta G_{cat}$ (from k_{cat} values) yields qualitatively the same results (not shown) as shown in Tables I and II and Figures 1 and 2. Thus, deviations from simple additivity are not systematically found in either the energetics to form the E-S complex or those to reach E-S‡.

Additive Effects on Protein–Protein Interactions

The first clear examples of additive binding effects caused by amino acid replacements in proteins were reported by Laskowski et al. (1983) and reviewed by others (Ackers & Smith, 1985; Horovitz & Rigbi, 1985). One hundred natural variants of a proteinase inhibitor, the ovomucoid third domain, have been isolated and sequenced from the eggs of different bird species (Empie & Laskowski, 1982; Laskowski et al., 1987). This is a nested set of proteins because for any one of these avian inhibitors there is a close relative containing only one or a few amino acid substitutions. Moreover, the association constants (K_a) of these inhibitors with a variety of serine proteinases vary over an enormous range (10^9-fold). Laskowski et al. (1983, 1989) have shown that the effect of a given residue replacement on K_a is about the same irrespective of the inhibitor scaffold the replacement is made in.

In addition to ovomucoid, four additivity examples have been constructed from natural variants at the subunit interface of tetrameric hemoglobin (Ackers & Smith, 1985). Three additivity examples have been analyzed for interactions of hGH with its receptor (B. C. Cunningham and J. A. Wells, unpublished results) and one example for association of synthetic variants of the RNase S peptide with RNase S protein (Mitchinson & Baldwin, 1986). The entirety of this data set is not tabulated because much on the ovomucoid inhibitors and hGH is unpublished. Nonetheless, these researchers were kind enough to provide their data formatted so it could be plotted collectively in Figure 4. These data consist of 91 additivity examples (80 in ovomucoids alone), representing 22 multiple mutants across four different proteins, and span a wide range of change in binding free energy (−10 to +7

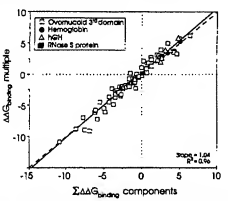


FIGURE 4: Plot showing the sum of changes in free energies of binding at protein–protein interfaces for component mutants versus the corresponding multiple mutant. Data represent interactions between ovomucoid third domain and various serine proteases (□) (R. Wynn and M. Laskowski, personal communication), regulatory interface of $\alpha_2\beta_2$ hemoglobin (●) (Ackers & Smith, 1985), hGH and its receptor (stippled △) (B. Cunningham and J. Wells, personal communication), and RNase S peptide and S protein (■) (Mitchinson & Baldwin, 1986). The dashed line represents a line of unity slope, and the solid line is the best fit.

kcal/mol). The plot shows a very strong linear correlation (R^2 = 0.96) with a slope near unity. Although the data for the ovomucoid were not sorted to evaluate changes at intramolecular contact sites, most are not expected to be in contact, and all of the other examples represent noncontact sites. Thus, the large data base derived from natural variants of ovomucoid third domain, as well as a smaller number of examples from several other proteins, indicates that multiple mutations at protein–protein interfaces commonly produce simple additive effects.

Additive Effects in DNA–Protein Interactions

One of the clear advantages in analyzing DNA–protein interactions is the ability to apply powerful selections that make analysis by random mutational studies feasible. Additivity in DNA–protein interactions was first demonstrated by reversion analysis of λ repressor (Nelson & Sauer, 1985). A mutation that decreased the binding affinity for the λ operator site (K4Q) was reverted by mutations at several second sites (E34K, G46S, and E83K). When these second-site reversants were introduced into wild-type λ repressor, they caused increases in affinity similar to those observed in the first-site suppressor mutant (K4Q).

Functional independence for mutations at DNA–protein contacts has been demonstrated by additive effects for mutants of CAP (catabolite gene activator protein) and its operator sequence (Ebright et al., 1987) as well as *lac* repressor and its corresponding operator sequence (Ebright, 1986). Simple additivity of mutational effects in the operator sequences for Cro repressor (Takeda et al., 1989) and λ repressor (Sarai & Takeda, 1989) has been most systematically demonstrated. Simple additivity has also been reported for multiple mutations in the *lac* repressor (Lehming et al., 1990). In fact, simple additivity is so predictable in DNA–protein interactions that the observation of complex additivity has been used to predict specific DNA–protein contacts in the *lac* repressor–operator complex (Ebright, 1986).

Additive Effects on Protein Stability

The first systematic analysis of additive effects of site-specific mutations on protein stability was reported by Shortle and Meeker (1986). Five multiple mutants in staphylococcal nuclease were constructed from a group of random single mutants that were screened initially for their ability to affect the stability of the enzyme in vivo. The component mutants do not make direct contact with each other in the multiple mutants. Generally, these variants exhibit nearly additive effects except for the double mutant V66L,G88V (Table IV). In addition to those of staphylococcal nuclease, additive effects on the $\Delta\Delta G_{unfolding}$ (assayed by reversible denaturation) have also been determined for the N-terminal domain of λ repressor (one example; Hecht et al., 1986), the α-subunit of *E. coli* Trp synthetase (one example; Hurle et al., 1986), T4 lysozyme (six examples; Wetzel et al., 1988), the gene V product of bacteriophage f1 (one example; Sandberg & Terwilliger, 1989),

Table IV: Comparison of Sums of $\Delta\Delta G_{unfolding}$ from Component Mutants vs the Multiple Mutant

assay	$\Delta\Delta G_{unfolding}$: component mutants	sum	multiple mutant
	Staphylococcal Nuclease		
	V66L + G79S[a]		
GuHCl	−0.2 −2.6	−2.8	−3.3
urea	+0.2 −2.9	−2.7	−3.6
	V66L + G88V		
GuHCl	−0.2 −1.0	−1.2	−2.1
urea	+0.2 −0.9	−0.7	−1.4
	I18M + A69T		
GuHCl	−0.6 −2.7	−3.3	−2.8
urea	−0.7 −2.9	−3.6	−3.8
	I18M + A90S		
GuHCl	−0.6 −1.4	−2.0	−2.2
urea	−0.7 −1.4	−2.1	−2.2
	V66L + G79S + G88V		
GuHCl	−0.2 −2.6 −1.0	−3.8	−3.0
urea	+0.2 −2.9 −0.9	−3.6	−3.4
	N-Terminal Domain of λ Repressor		
	G46A + G48A[b]		
thermal melt	+0.7 +0.9	+1.6	+1.1
	T4 Lysozyme		
	I3C + C54V[c]		
thermal melt	+1.2 −0.7	+0.5	+0.4
	I3C + C54T		
thermal melt	+1.2 +0.3	+1.5	+1.5
	I3C + C54T + R96H		
thermal melt	+1.2 +0.3 −2.8	−1.3	−2.5
	I3C,C54T + R96H		
thermal melt	+1.5 −2.8	−1.3	−2.5
	I3C + C54T + A146T		
thermal melt	+1.2 +0.3 −1.5	0	−0.5
	I3C,C54T + A146T		
thermal melt	+1.5 −1.5	0	−0.5
	Bacteriophage f1 Gene V		
	V35I + I47V[d]		
GuHCl	−0.4 −2.4	−2.8	−2.9
	Kringle-2 of tPA		
	H64Y + R68G[e]		
thermal melt	+2.9 +0.7	+3.6	+3.4
	Turkey Ovomucoid Third Domain		
	G32A + N28S[f]		
thermal melt	+0.8 −0.5	+0.3	+0.2
	Y20H + N45-CHO		
thermal melt	−0.8 +0.3	−0.5	−0.6
	α Subunit of *E. coli* Trp Synthetase		
	Y175C + G211E[g]		
GuHCl	−0.1 +0.3	+0.2	−1.3

[a] Shortle and Meeker (1986). [b] Hecht et al. (1986). [c] Wetzel et al. (1988). [d] Sandberg and Terwilliger (1989). [e] R. Kelley, personal communication. [f] Otlewski and Laskowski (1990). N45-CHO refers to a glycosylation of Asn45. [g] Hurle et al. (1986).

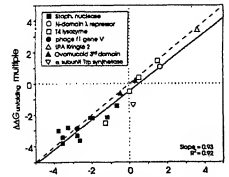


FIGURE 5: Plot showing sum of changes in free energy of unfolding of component mutants and resulting multiple mutant. Data are taken from Table IV and represent staphylococcal nuclease (■), N-terminal domain of λ repressor (○), T4 lysozyme (□), bacteriophage f1 gene V product (●), Kringle-2 domain of tissue plasminogen activator (▲), turkey ovomucoid third domain (△), and the α-subunit of Trp synthetase (▽). The dashed line represents a theoretical line of unity slope, and the solid line represents the best fit.

natural variants of ovomucoid third domain (two examples; Otlewski & Laskowski, 1990), and the Kringle-2 domain of human tissue plasminogen activator (t-PA) (one example; R. Kelley, personal communication).

Collectively, this data set gives a high linear correlation (R^2 = 0.94) and slope near unity (Figure 5). The generally simple additive behavior is somewhat surprising given the highly cooperative nature of protein folding. There are discrepancies in some of the additivity examples besides the staphylococcal nuclease mutant (V66L,G88V). For example, the 1.5 kcal/mol discrepancy for the Y175C,G271E double mutant in Trp synthetase (Table IV) is proposed to result from the fact that these residues are in direct contact (Hurle et al., 1986). Furthermore, proximity effects may account for the large differences between the sum of the component mutants and the multiple mutants for the α-helical double glycine mutant G46A,G48A in λ repressor (Hecht et al., 1986), and when combining R96H with the C2–C97 disulfide mutant in T4 lysozyme (Wetzel et al., 1988). In contrast, an exchange of two side chains that contact one another (V35I and I47V) in the hydrophobic core of the gene V product of f1 phage produced simple additive effects (Sandberg & Terwilliger, 1989; Table IV). It should be noted that this data base exhibiting simple additivity may be biased for single mutants that stably fold, because severely unstable proteins are more difficult to express.

By analogy to transition-state binding effects, one can certainly imagine instances where the stabilizing effects of mutations should reach a plateau. For example, denaturation at high temperatures can become controlled by a chemical step such as deamidation (Ahern et al., 1987), so that additional mutants that stabilize the folded form of the protein may be irrelevant. Another obvious example where complex additivity can be observed in protein stability is the stabilizing effect of disulfide bonds and noncovalent intramolecular contacts that require interactions between two or more residues. In these cases, the stabilizing interaction between two side chains can be broken with only one mutation.

Applications of Additivity in Rational Protein Design

A strategy of additive mutagenesis, where a series of single mutants each making a small improvement in function are combined, is one of the most powerful tools in designing functional properties in proteins. This approach has been remarkably successful in stabilizing proteins to irreversible inactivation, such as λ repressor (Hecht et al., 1986), subtilisin (Bryan et al., 1987; Cunningham & Wells, 1987; Pantoliano et al., 1989), kanamycin nucleotidyltransferase (Liao et al., 1986; Matsumura, 1986), neutral protease (Imanaka et al., 1986), and T4 lysozyme (Wetzel et al., 1988; Matsumara et al., 1989). This strategy has been applied to enhancing the catalytic efficiency of a weakly active variant of subtilisin (Carter et al., 1989), engineering the substrate specificity of subtilisin (Wells et al., 1987a,b; Russell & Fersht, 1987) and the coenzyme specificity of glutathione reductase (Scrutton et al., 1990), designing protease inhibitors with exquisite protease specificity (Laskowski et al., 1989), and recruiting human prolactin to bind to the hGH receptor (Cunningham et al., 1990). In addition, additivity principles have been used to engineer the pH profile of subtilisin (Russell & Fersht, 1987) and to design the affinity and specificity of λ repressor (Nelson & Sauer, 1985).

For this approach to work does not require that all the component mutants act in a simply additive manner but just that their effects accumulate. For example, despite the complex additivity of effects in the catalytic triad of subtilisin, there are mutagenic pathways that are energetically cumulative for installing the triad (Carter & Wells, 1988; Wells et al., 1987c). Starting with the triple mutant S221A,H64A,D32A, there is a progressive enhancement for installing Ser221 (~1.1 kcal/mol), then His64 (~1.0 kcal/mol), and finally Asp32 (~6.5 kcal/mol). Another cumulative pathway of Ser221, then Asp32, and finally His64 is possible if the Ser221,Asp32 intermediate were to use HisP2 substrates (Carter & Wells, 1987). Elaborating such cumulative pathways is important for understanding how a catalytic apparatus may have evolved and is practically useful for considering how to install such catalytic machinery into weakly active catalytic antibodies.

Conclusions

In the majority of cases, combination of mutations that affect substrate or transition-state binding, protein–protein interactions, DNA–protein recognition, or protein stability exhibits simple additivity. Simple additivity is commonly observed for distant mutations at rigid molecular interfaces such as in protein–protein and DNA–protein interactions, where the mutations are unlikely to alter grossly the structure or mode of binding.

Large deviations from simple additivity can occur when the sites of mutations strongly interact with one another (by making direct contact or indirectly through electrostatic interactions or large structural perturbations) and/or when both sites function cooperatively (as for the catalytic triad and oxyanion binding site of subtilisin). Changes at sites that can contact each other do not always lead to complex additivity; this may reflect relatively weak interactions between the two sites or indicate that the interactions are compensatory and appear to be weak.

It is important to point out the magnitude of errors in predicting the free energy effect in the multiple mutant from the component single mutants. Generally, for those cases exhibiting simple additivity (Figures 1, 4, and 5), the discrepancy in free energy between the sums of the components and multiple mutants is about ±25%. Part of this is the result of compounding errors when summing the single mutants, and the rest is presumably due to weak interaction terms. Nonetheless, this means that if the total free energy change is about 3 kcal/mol, the change in the equilibrium constant

(related by $K_{xy}/K_{wz} = 10^{-3/RT} \approx 155$) will often be off by a factor of 4. Thus, while the free energy effects accumulate, significant deviations will occur in predicting the final equilibrium constants when component mutants contribute a large free energy term.

Simple additivity reflects the modularity of component amino acids in protein function. This results from the fact that the perturbations in energetics and structure resulting from most mutations are highly localized. In the past six years, an additive mutagenesis strategy has been extremely effective in engineering proteins—of course, nature has been using this strategy much longer.

ACKNOWLEDGMENTS

I am grateful to Dr. Michael Laskowski and Rich Wynn for providing their data prior to publication on ovomucoid third domain and similarly to Brian Cunningham, Paul Carter, and Robert Kelley for making available their unpublished data. I am indebted for useful discussions with Drs. Michael Laskowski, Paul Carter, Jack Kirsch, and Tony Kossiakoff and many of my colleagues at Genentech and to Drs. Richard Ebright and William Jencks and those above for critical reading of the manuscript.

Registry No. RNase, 9001-99-4; tyrosyl-tRNA synthetase, 9023-45-4; trypsin, 9002-07-7; dihydrofolate reductase, 9002-03-3; subtilisin BPN′, 9014-01-1; glutathione reductase, 9001-48-3; staphylococcal nuclease, 9013-53-0; lysozyme, 9001-63-2; plasminogen activator, 105913-11-9; tryptophan synthetase, 9014-52-2.

REFERENCES

Ackers, G. K., & Smith, F. R. (1985) *Annu. Rev. Biochem. 54*, 597–629.

Ahern, T. J., Casal, J. I., Petsko, G. A., & Klibanov, A. M. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 675–679.

Alber, T., Dao-pin, S., Wilson, K., Wozniak, J. A., Cook, S. P., & Matthews, B. W. (1987) *Nature 330*, 41–46.

Albery, W. J., & Knowles, J. R. (1976) *Biochemistry 15*, 5631–5640.

Ardelt, W., & Laskowski, M., Jr. (1990) *J. Mol. Biol.* (submitted for publication).

Bott, R., Ultsch, M., Wells, J., Powers, D., Burdick, D., Struble, M., Burnier, J., Estell, D., Miller, J., Graycar, T., Adams, R., & Power, S. (1987) *ACS Symposium Series 334* (LeBaron, H. M., Mumma, R. O., Honeycutt, R. C., & Deusing, J. H., Eds.) pp 139–147, American Chemical Society, Washington, DC.

Brown, K. A., Brick, P., & Blow, D. M. (1987) *Nature 326*, 416–418.

Bryan, P. N., Rollence, M.-L., Pantoliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J., & Poulos, T. L. (1987) *Proteins: Struct., Funct., Genet. 1*, 326–334.

Carter, P., & Wells, J. A. (1987) *Science 237*, 394.

Carter, P., & Wells, J. A. (1988) *Nature 332*, 564–568.

Carter, P., & Wells, J. A. (1990) *Proteins: Struct., Funct., Genet.* (in press).

Carter, P., Nilsson, B., Burnier, J. P., Burdick, D., & Wells, J. A. (1989) *Proteins: Struct., Funct., Genet. 6*, 240–248.

Carter, P. J., Winter, G., Wilkinson, A. J., & Fersht, A. R. (1984) *Cell 38*, 835–840.

Chothia, C., & Lesk, A. (1986) *EMBO J. 5*, 823–826.

Craik, C. S., Largman, C., Fletcher, T., Roczniak, S., Barr, P. J., Fletterick, R., & Rutter, W. J. (1985) *Science 228*, 291–297.

Cunningham, B. C., & Wells, J. A. (1987) *Protein Eng. 1*, 319–325.

Cunningham, B. C., Henner, D. J., & Wells, J. A. (1990) *Science 247*, 1461–1465.

Ebright, R. H. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 303–307.

Ebright, R. H., Kolb, A., Buc, H., Kunkel, T. A., Krakow, J. S., & Beckwith, J. (1987) *Proc. Natl. Acad. sci. U.S.A. 84*, 6083–6087.

Empie, M. W., & Laskowski, M., Jr. (1982) *Biochemistry 21*, 2274–2284.

Fersht, A. (1985) in *Enzyme Structure and Mechanism*, 2nd ed., Chapters 3, 12, and 13, W. H. Freeman and Co., New York.

Fersht, A. R. (1987) *Biochemistry 26*, 8031–8037.

Fersht, A. R., Wilkinson, A. J., Carter, P., & Winter, G. (1985) *Biochemistry 24*, 5858–5861.

Hecht, M. H., Sturtevant, J. M., & Sauer, R. T. (1986) *Proteins: Struct., Funct., Genet. 1* 43–46.

Horowitz, A., & Rigbi, M. (1985) *J. Theor. Biol. 116*, 149–159.

Howell, E. E., Villafranca, J. E., Warren, M. S., Oatley, S. J., & Kraut, J. (1986) *Science 231*, 1123–1128.

Hurle, M. R., Tweedy, N. B., & Matthews, C. R. (1986) *Biochemistry 25*, 6356–6360.

Imanaka, T., Shibazaki, M., & Takagi, M. (1986) *Nature 324*, 695–697.

Jencks, W. P. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4046–4050.

Jones, M. D., Lowe, D. M., Borgford, T., & Fersht, A. R. (1986) *Biochemistry 25*, 1887–1891.

Katz, B. A., & Kossiakoff, A. (1986) *J. Biol. Chem. 261*, 15480–15485.

Laskowski, M., Jr., Tashiro, M., Empie, M. W., Park, S. J., Kato, I., Ardelt, W., & Wieczorek, M. (1983) in *Protease Inhibitors: Medical and Biological Aspects* (Katunuma, N., Ed.) pp 55–68, Japan Scientific Societies Press, Tokyo, Japan.

Laskowski, M., Jr., Kato, I., Ardelt, W., Cook, J., Denton, A., Empie, M. W., Kohr, W. J., Park, S. J., Parks, K., Schatzley, B. L., Schoenberger, O. L., Tashiro, M., Vichot, G., Whatley, H. E., Wieczorek, A., & Wieczorek, M. (1987) *Biochemistry 26*, 202–221.

Laskowski, M., Jr., Park, S. J., Tashiro, M., & Wynn, R. (1989) in *Protein Recognition of Immobilized Ligands*, UCLA Symposia on Molecular and Cellular Biology (Hutchens, T. W., Ed.) Vol. 80, pp 149–160, A. R. Liss, New York.

Leatherbarrow, R. J., Fersht, A. R., & Winter, G. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 7840–7844.

Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B., & Muller-Hill, B. (1990) *EMBO J. 9*, 615–621.

Liao, H., McKenzie, T., & Hageman, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 83*, 576–580.

Lowe, D. M., Fersht, A. R., Wilkinson, A. J., Carter, P., & Winter, G. (1985) *Biochemistry 24*, 5106–5109.

Matsumura, M., Yasumura, S., & Aiba, S. (1986) *Nature 323*, 356–358.

Matsumura, M., Signor, G., & Matthews, B. W. (1989) *Nature 342*, 291–294.

Matthews, B. W. (1987) *Biochemistry 26*, 6885–6888.

Mayer, R. J., Chen, J.-T., Taira, K., Fierke, C. A., & Benkovic, S. J. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 7718–7720.

Mitchinson, C., & Baldwin, R. L. (1986) *Proteins: Struct., Funct., Genet. 1*, 23–33.

Nelson, H. C. M., & Sauer, R. T. (1985) *Cell 42*, 549–558.
Otlewski, J., & Laskowski, M., Jr. (1990) (submitted for publication).
Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989) *Biochemistry 28*, 7205–7213.
Russell, A. J., & Fersht, A. R. (1987) *Nature 328*, 496–500.
Sandberg, W. S., & Terwilliger, T. C. (1989) *Science 245*, 54–57.
Sarai. A., & Takeda, Y. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 6513–6517.
Schechter, I., & Berger, A. (1967) *Biochem. Biophys. Res. Commun. 27*, 157–162.
Scrutton, N. S., Berry, A., & Perham, R. N. (1990) *Nature 343*, 38–43.
Shortle, D., & Meeker, A. K. (1986) *Proteins: Struct., Funct., Genet. 1*, 81–89.
Takeda, Y., Sarai, A., & Rivera, V. M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 439–443.
Wells, J. A., & Estell, D. A. (1988) *Trends Biochem. Sci. 13*, 291–297.
Wells, J. A., Cunningham, B. C., Graycar, T. P., & Estell, D. A. (1987a) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5167–5171.
Wells, J. A., Powers, D. B., Bott, R. R., Graycar, T. P., & Estell, D. A. (1987b) *Proc. Natl. Acad. Sci. U.S.A. 84*, 1219–1223.
Wells, J. A., Cunningham, B. C., Graycar, T. P., Estell, D. A., & Carter, P. (1987c) *Cold Spring Harbor Symp. Quant. Biol. 52*, 647–652.
Wetzel, R., Perry, L. J., Baase, W. A., & Becktel, W. J. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 401–405.
Wilde, J. A., Bolton, P. H., Dell'Acqua, M., Hibler, D. W., Pourmotabbed, T., & Gerit, J. A. (1988) *Biochemistry 27*, 4127–4132.
Wilkinson, A. J., Fersht, A. R., Blow, D. M., & Winter, G. (1983) *Biochemistry 22*, 3581–3586.
Wilkinson, A. J., Fersht, A. R., Blow, D. M., Carter, P., & Winter, G. (1984) *Nature 307*, 187–188.

Accelerated Publications

Role of Tyrosine M210 in the Initial Charge Separation of Reaction Centers of *Rhodobacter sphaeroides*†

Ulrich Finkele, Christoph Lauterwasser, and Wolfgang Zinth
Physik Department der Technischen Universität, D-8000 München 2, FRG
Kevin A. Gray and Dieter Oesterhelt*
Max-Planck-Institut für Biochemie, D-8033 Martinsried, FRG
Received May 22, 1990; Revised Manuscript Received July 12, 1990

ABSTRACT: Femtosecond spectroscopy was used in combination with site-directed mutagenesis to study the influence of tyrosine M210 (YM210) on the primary electron transfer in the reaction center of *Rhodobacter sphaeroides*. The exchange of YM210 to phenylalanine caused the time constant of primary electron transfer to increase from 3.5 ± 0.4 ps to 16 ± 6 ps while the exchange to leucine increased the time constant even more to 22 ± 8 ps. The results suggest that tyrosine M210 is important for the fast rate of the primary electron transfer.

The primary photochemical event during photosynthesis of bacteriochlorophyll- (Bchl-) containing organisms is a light-induced charge separation within a transmembrane protein complex called the reaction center (RC). The crystal structures of RC's from *Rhodopseudomonas* (*Rps.*) *viridis* and *Rhodobacter* (*Rb.*) *sphaeroides* have been solved to high resolution [reviewed in Deisenhofer and Michel (1989), Chang et al. (1986), Tiede et al. (1988), and Rees et al. (1989)]. The RC from *Rb. sphaeroides* contains three protein subunits referred to as L, M, and H, according to their respective mobilities in SDS–polyacrylamide gels. Associated with the L and M subunits are the cofactors, consisting of four Bchl *a*, two bacteriopheophytin (Bph) *a*, one atom of non-heme ferrous iron, two quinones (Q_A and Q_B), and in some species one carotenoid [reviewed in Parson (1987) and Feher et al. (1989)]. The cofactors are arranged in two branches (Figure 1) with an approximate C_2 axis of symmetry. The kinetic data support a model in which the primary electron transfer proceeds after light absorption by the primary donor [a special pair of Bchl referred to as P; reviewed in Kirmaier and Holten (1987)]. The absorption of light generates the excited electronic state P*, which has a lifetime of approximately 3 ps. An electron is transferred from P along only one branch (the so-called A-branch). It is generally accepted that after approximately 3 ps the electron arrives at the Bph on the A-side (H_A) and after 220 ps it reaches Q_A. The role of the accessory Bchl located between P and H_A (referred to as B_A) has not been definitely assigned. Recently, we have shown that at room temperature an additional kinetic (τ = 0.9 ps) component is detectable (Holzapfel et al., 1989). The spectral properties and the kinetic constants lead to the conclusion that the corresponding intermediate is the radical pair $P^+B_A^-$ (Holzapfel et al., 1990).

Additional intriguing points concerning the process of

† Financial support was from the Deutsche Forschungsgemeinschaft, SFB 143.
* To whom correspondence should be addressed.

0006-2960/90/0429-8517$02.50/0

TECHVIEW: GENOMICS

The Babel of Bioinformatics

Teresa K. Attwood

The sequencing of entire genomes is a major achievement, but the meaning of the mass of accumulated data is only just beginning to be unraveled. At first sight, the task appears straightforward: locate the genes and translate the coding regions to establish their protein products; perform similarity searches to establish relationships with previously characterized sequences and assign function by evolutionary inference; and rationalize the function in structural terms using known or model-derived structures. Given the quantity of data, the procedures should be automated as much as possible.

The reality, of course, is not so simple. Attempts to decipher the clues latent in genomic data are hampered because current methods to predict genes in uncharacterized DNA are unreliable (and it is not always clear what we mean by "gene"); it is presumptuous to make functional assignments merely on the basis of some degree of similarity between sequences (and it is not always clear what we mean by "function"); very few structures are known compared with the number of sequences, and structure prediction methods are unreliable (and knowing structure does not inherently tell us function); the degree of automation that has been used of necessity, with its imperfect tools and protocols, has led to the accumulation of much database misinformation; and the terminology has been imprecise, muddying perceptions of what can realistically be achieved. Given these problems, what is the state of the art in sequence-structure-function bioinformatics?

Gene prediction

Information used to predict genes includes signals in the sequence, content statistics, and similarity to known genes. In a recent test of gene detection tools as part of the *Drosophila* genome, the majority of these "gene finders" identified 95% of coding nucleotides, but intron/exon structures were correctly predicted for only about 40% of genes. Ten different methods failed to find between 5% and 95% of genes, and incorrectly identified up to 55% (*1*). But probably the most sobering evidence of the frailty of gene prediction methods is the uncertainty in the number of genes in the human genome, with current estimates ranging from 27,462 to 312,278. The methods used to arrive at these numbers each involve different approximations and extrapolations. Nevertheless, it is disturbing that the different analytical approaches should yield such disparate results.

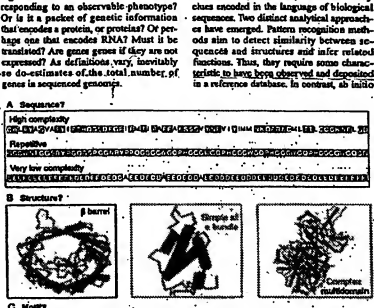
What is a gene?

Perhaps the biggest obstacle to accurate gene counting is that even the definition of a gene is unclear. Is it a heritable unit corresponding to an observable phenotype? Or is it a packet of genetic information that encodes a protein, or proteins? Or perhaps one that encodes RNA? Must it be translated? Are genes genes if they are not expressed? As definitions vary, inevitably so do estimates of the total number of genes in sequenced genomes.

The sequence-structure imbalance

To date, more than 540,000 protein sequences have been deposited in the nonredundant database maintained by the National Center for Biotechnology Information (NCBI), and millions of expressed sequence tags (ESTs), which are partial sequences of clones that are often error prone, are housed in public and proprietary repositories. These numbers will snowball with the fruition of further genome projects. By contrast, the number of unique protein structures is still less than 2000. Of course, we do not know how many unique sequences there are; nevertheless, it is clear that there is a dearth of structural information.

Given this sequence-structure imbalance, it is imperative that we focus on deciphering the structural, functional, and evolutionary clues encoded in the languages of biological sequences. Two distinct analytical approaches have emerged. Pattern recognition methods aim to detect similarity between sequences and structures and infer related functions. Thus, they require some characteristic to have been observed and deposited in a reference database. In contrast, ab initio



Levels of complexity. Looking at a sequence (**A**) or a fold (**B**) in isolation, we can say little about its function. Only when we look at sequences or structures together do the patterns of conservation that emerge (motifs) begin to provide functional clues. For example, the above motifs (**C**) may suggest roles in calcium binding, nucleotide binding, and membrane anchoring. We can think of folds as providing different scaffolds, which can be decorated in different ways by different sequences to confer different functions. Knowing both the fold and the function allows us to rationalize the mechanism by which the structure affects its function at the molecular level.

The author is at the School of Biological Sciences, University of Manchester, Manchester, UK. E-mail: attwood@bioinf.man.ac.uk

prediction methods deduce structure directly from sequence. The approaches are quite different and should not be confused. Their levels of success also differ markedly.

Function prediction through pattern recognition

Tools for similarity searching are standard components of the sequence annotator's armory. Sequence similarity programs rapidly seek pairwise similarities in large sequence repositories or search for conserved patterns in gene family databases (*2–5*). Gene family databases allow more specific functional diagnoses to be made than is possible by pairwise searching. They are based on the principle that related sequences can be aligned to find regions (motifs) that show little variation. These motifs usually reflect some vital structural or functional role (see the figure), and they can be used to derive diagnostic family signatures. Sequences can then be searched against databases of such signatures to see whether they can be assigned to known families. Gene family databases have recently been integrated to create a unified protein family resource (*6*), facilitating the inference of function by identifying homologous relationships.

The term "homology" is fundamental concept in bioinformatics; it often used incorrectly. Sequences are homologous if they are related by divergence from a common ancestor (*7*). Conversely, analogy relates to the acquisition of common structural or functional features via convergent evolution from unrelated ancestors. For example, β barrels occur in soluble serine proteases and integral membrane porins, but despite their common architecture, they share no sequence or functional similarity. Similarly, the enzymes chymotrypsin and subtilisin share groups of catalytic residues with almost identical spatial geometries, but they have no other sequence or structural similarities. Homology is not a measure of similarity, but rather an absolute statement that sequences have a divergent rather than a convergent relationship. This is not just a semantic issue because confusion over the term obscures evolutionary relationships. In comparing structures, the same arguments apply. Structures may be similar, but common evolutionary origin remains a hypothesis until supported by other evidence; the hypothesis may be correct or mistaken, but the similarity is a fact (*8*).

Among homologous sequences, we can distinguish orthologs (proteins that usually perform the same function in different species) and paralogs (proteins that perform different but related functions within one organism). Orthologs allow investigation of cross-species relationships, whereas paralogs, which arise via gene duplication events, shed light on underlying evolutionary mechanisms because the duplicated genes follow separate evolutionary pathways and new specificities evolve through variation and adaptation. Such complexity presents real challenges for bioinformatics. When analyzing a database search, it may be unclear how much functional annotation can be legitimately inherited by a query sequence, and whether the best match turned up by the search is the true ortholog or a paralog. This difficulty is the source of numerous annotation errors.

Further complications result from the domain and/or modular nature of many proteins. Modules are autonomous folding units that often function as protein building blocks, forming multiple combinations of the same module or mosaics of different modules. They can confer a variety of functions on the parent protein. If the best hit in a database search is related to a single domain or module, it is unlikely that the function annotation can be propagated from the parent protein to the query sequence.

In using modules to confer different functionalities, Nature uses old material to create new systems. The complexity of such systems poses important problems for computational approaches because the properties of a system can be explained by but not deduced from those of its components (*9*, *10*). The presence of a module tells little of the function of the complete system; knowing most components of a mosaic does not allow us easily to predict a missing one, and modules in different proteins do not always perform the same function.

Many other factors also complicate function assignment: gene functions may be redundant, nonorthologous displacement can replace genes with unrelated but functionally analogous genes, horizontal gene transfer can introduce genes from different phylogenetic lineages, and lineage-specific gene loss can eliminate ancestral genes. Thus, genomes harbor many obstacles to reliable function assignment.

What is function?

Protein function is context-dependent. Vagueness in using the term has yielded confusing database annotations. It is currently used to refer variously to biochemical activities, biological goals, and cellular structure; for example, the function of actin might be described as "ATPase" or "constituent of the cytoskeleton." In an attempt to introduce rigor into the field and better relate biological reality, independent ontologies such as the Gene Ontology (*11*) are under development that aim to define more explicitly the relationships between gene products and biological processes, molecular functions, and cellular components.

Structure prediction and fold recognition

We have seen that definitions of "genes" differ, making it difficult to count genes accurately, and that our concepts of "function" differ, making function assignment tricky. It would seem, however, that we can agree on what structures are. They are tangible, measurable things, so should we not be able to predict them reliably?

Structure prediction methods range from computationally intensive strategies that simulate the physical and chemical forces involved in protein folding to knowledge-based approaches that use information from structure databases to build models. Yet the problem of predicting protein structure remains unsolved: knowledge-based techniques typically produce low-resolution models, and no current method yields reliable predictions for remote homologs (*12*). For small proteins, ab initio methods generate models with substantial segments that resemble the correct fold, but results deteriorate beyond ~100 residues. Today, knowledge-based methods, especially those that combine information from different approaches, give best results (*13*). The most successful modeling and fold recognition studies have balanced better algorithms with appropriate levels of manual intervention (*14*).

Prediction methods do not work well because we do not fully understand how the primary structure of a protein determines its tertiary structure. Structural genomics projects will gradually lessen our reliance on prediction, because they aim to provide experimental structures or models for every protein in all completed genomes (although membrane protein structures will be difficult to obtain because they are difficult to crystallize). We must keep in mind, however, that structure alone will not inherently tell us function (see the figure). For example, determining the structure of a hypothetical protein and discovering that it binds ATP (*15*) may shed light on possible aspects of its functionality, but such information does not reveal its specific biological function.

What is structure?

In this context of fold recognition and prediction, it is important to be precise about what we mean by "structure." For example, is a prediction a "good" prediction if it correctly reproduces all atomic positions, the topology (connectivity of secondary structures), the architecture (gross arrangement of secondary structures), or merely the structural class (mainly α, mainly β, etc.)? Where does a "reasonably good" prediction fall in this hierarchy, and what level of structural detail does a "tough near miss" (*16*) reveal? Using such imprecise words hinders comprehension, making it difficult to evaluate what a good prediction really is.

Outlook

In "predicting" genes, protein functions, and structures, it is helpful to define our terms precisely and be honest about our achievements. Otherwise, we will continue to be baffled by paradoxical new prediction methods that yield >80% error rates. Gene identification, structure prediction, and functional inference are nontrivial computational tasks, but with the relentless accumulation of sequence data, improvements continue to be made in all areas.

Nature functions by integration, and the adoption of a more holistic view of complex biological systems is an essential next step for bioinformatics. To get the most from genomic data, we need to take account of information on the regulation of gene expression, metabolic pathways, and signaling cascades. Proteins do not work in isolation but are involved in interrelated networks. Unraveling these networks and their interactions will be vital to our understanding of normal and pathologic cell development, and will help us create an integrated mapping between genotype and phenotype.

Genomics-based drug discovery is heavily dependent on accurate functional annotation. Toward this end, bioinformatics will need to deliver highly integrated, interoperable databases (and data "warehouses") that allow the user to reason over disparate data sources and ultimately enable knowledge-based inference and innovation. The more genome annotation is automated, the greater will be the need for collaborations between software developers, annotators, and experimentalists. And the more data we have to handle, the more rigorous we must be in our thinking. Only if we are to make sense of the complexities. Sequence-structure-function bioinformatics does not yet yield all the answers, but a future holistic approach should help fuse today's glimmerings of knowledge into a new dawn of understanding.

References and Notes

1. M. O. Dayhoff et al., *Comput. Biol.* **3**, 483 (2000).
2. E. Huberman et al., *Nucleic Acids Res.* **27**, 215 (1999).
3. T. K. Attwood et al., *Nucleic Acids Res.* **28**, 225 (2000).
4. A. Bateman et al., *Nucleic Acids Res.* **28**, 263 (2000).
5. J. Hardwell et al., *Nucleic Acids Res.* **28**, 228 (2000).
6. R. Apweiler et al., *Bioinformatics*, in press.
7. W. M. Rosh, *Syst. Zool.* **15**, 99 (1970).
8. G. G. Simpson et al., *Cell* **90**, 601 (1997).
9. F. Jacob, *Science* **196**, 1161 (1977).
10. L. Cobb et al., *Curr. Opin. Genet. Dev.* **7**, 648 (1997).
11. M. Ashburner et al., *Nature Genet.* **25**, 25 (2000).
12. S. Rost and L. O'Donoghue, *Comput. Appl. Biosci.* **13**, 345 (1997).
13. A. R. Panchenko et al., *J. Mol. Biol.* **296**, 1391 (2000).
14. M. J. E. Sternberg et al., *Curr. Opin. Struct. Biol.* **8**, 349 (1998).
15. T. L. Zarembinski et al., *Proc. Natl. Acad. Sci. U.S.A.* **95**, 15189 (1998).
16. E.A. Ostrovsky et al., *Comput. Chem.* **24**, 499 (2000).
17. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

SCIENCE'S COMPASS

TECHSIGHTING

SOFTWARE

Conquering by Dividing

The average personal computer spends much less than half a day actually performing useful computations. Many users, concerned about the vulnerability of expensive electronic components to the constant cycling of the power on and off, leave their systems on continuously. It is staggering to imagine the enormous, untapped computing resources of several million PCs left running unattended. One popular approach to tapping this computing power is the Search for Extra-Terrestrial Intelligence (SETI) project (*1*), which breaks giant computing problems into pieces that can be solved on personal computers in their spare time.

Popular Power, Inc. is a company offering a new twist on this theme. Like SETI, a company computer feeds pieces of large computing problems to networked personal computers via their software program, Popular Power Worker, for idle-time operation. Popular Power's approach differs, however, in providing a variety of computing problems to work on. Those include nonprofit projects with no financial incentive to the personal computer owner, as well as commercial jobs that will eventually pay users for tasks performed on their machines.

The current version of the Popular Power Worker runs only on Windows and Linux systems and is officially in pre-release form. The preliminary status of the software is readily apparent; numerous bugs, frequent crashes, and difficulties in installation plague the program currently. If information at the company Web site is accurate, personal computer owners interested in Popular Power's computing model may find dealing with the problems of the early release worth their while. Users of the pre-release software are promised priority of access to commercial computing jobs after the official version is released. Popular Power Worker can be downloaded for free from the company's Web site, and it installs as a screen saver, which starts the program running when it becomes active. Future versions of the program for Macintosh and Solaris systems are planned.

> **Popular Power Worker**
> Popular Power, Inc.
> San Francisco, CA
> Free
> www.popularpower.com

The benefits of the Popular Power scheme for distributed computing tasks do not accrue solely to the user whose computer is used. The flexible nature of Popular Power's design provides access for businesses, scientists, and anyone with massive computing projects to computing power that is potentially far greater than they would gain from a fixed piece of hardware. Personal computer users might be able to select which commercial job to run through Popular Power Worker depending on the return offered by the originating contractor. A key to the success of the computing model is likely to be the price Popular Power demands for acting as the interface between the computing project creators and the personal computer users.

In summary, the current version of Popular Power Worker is still in the testing phase and users may find the software unstable. Tech-savvy pre-2001 computer enthusiasts are best suited to test the current pre-release product. The remaining users are advised to wait at least for the official release of the software.

—KEVIN AHERN

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA. E-mail: kherst@bcc.orst.edu

References

1. J. Kahar, *Science* **282**, 839 (1998).

TECHSIGHTING

SOFTWARE

Eyes on the Skies

The orbital space above Earth contains an astonishing collection of man-made satellites. Tracking all of these objects is no small task. Liftoff is a NASA Web site that provides several software tools to locate, track, and identify Earth-orbiting satellites. At the Web site, three programs are available: J-Pass (identifies satellites passing overhead); J-Track (allows one to track orbiting objects); and J-Track 3D (allows one to view satellites orbiting Earth from a perspective far away in space). Each of these platform-independent applications is written in Java and is accessible from both Internet Explorer and Netscape

> **J-Track, J-Track 3D, and J-Pass**
> NASA
> Free
> http://liftoff.msfc.nasa.gov/realtime/JTrack/SpacecraftJ.html

Tech.Sight is published in the third issue of each month. Contributing editor: Kevin Ahern, Department of Biochemistry and Biophysics, Oregon State University. Send your comments by e-mail to techsight@aaas.org

From genes to protein structure and function: novel applications of computational approaches in the genomic era

Jeffrey Skolnick and Jacquelyn S. Fetrow

The genome-sequencing projects are providing a detailed 'parts list' of life. A key to comprehending this list is understanding the function of each gene and each protein at various levels. Sequence-based methods for function prediction are inadequate because of the multifunctional nature of proteins. However, just knowing the structure of the protein is also insufficient for prediction of multiple functional sites. Structural descriptors for protein functional sites are crucial for unlocking the secrets in both the sequence and structural-genomics projects.

Genome-sequencing projects are providing a detailed 'parts list' for life. Unfortunately, this list, a portion of which represents the amino acid sequence of all the proteins in a given genome, does not come with an instruction manual. That is, given the genome's sequences, one does not necessarily know straight away which regions encode proteins, which serve a regulatory role and which are responsible for the structure and replication of the DNA itself.

This is not unlike giving a child a list of parts necessary to create a working automobile. Without the necessary expertise, creating the final, working car from just the initial parts list is a nearly impossible task. Similarly, understanding how to create a complete, functioning cell given just the sequence of nucleotides found in an organism's genome is a complex problem.

What is a protein function?

After a genome is sequenced and its complete parts list determined, the next goal is to understand the function(s) of each part, including that of the proteins. What do we mean by protein function, the focus of this article?

Function has many meanings. At one level, the protein could be a globular protein, such as an enzyme, hormone or antibody, or it could be a structural or membrane-bound protein. Another level is its biochemical function, such as the chemical reaction and the substrate specificity of an enzyme. The regulatory molecules or cofactors that bind to a protein are also levels of biochemical function.

At the cellular level, the protein's function would involve its interaction with other macromolecules and the function and cellular location of such complexes. There is also the protein's physiological function; that is, in which metabolic pathway the protein is involved or what physiological role it performs in the organism. Finally, the phenotypic function is the role played by the protein in the total organism, which is observed by deleting or mutating the gene encoding the protein.

J. Skolnick (skolnick@danforthcenter.org) is at the Danforth Plant Science Center, Laboratory of Computational Genomics, 4041 Forest Park Avenue, St Louis, MO 63108, USA. J.S. Fetrow is at GeneFormatics, Suite 200, 5830 Oberlin Drive, San Diego, CA 92121-3754, USA.

Obviously, the complete characterization of protein function is difficult but efforts are under way at all levels[1–4], including cellular function[5,6]. In this article, however, we focus on identifying the biochemical function of a protein given its sequence, a problem that is amenable to molecular approaches.

Sequence-based approaches to function prediction

The sequence-to-function approach is the most commonly used function-prediction method. This robust field is well developed and, in the interest of space limitations, we will merely present a brief overview.

There are two main flavors of this approach: sequence alignment[7–9]; and sequence-motif methods such as Prosite[10], Blocks[11], Prints[12,13] and Emotif[14]. Both the alignment and the motif methods are powerful but a recent analysis has demonstrated their significant limitations[15], suggesting that these methods will increasingly fail as the protein-sequence databases become more diverse.

An extension of these approaches that combines protein-sequence with structural information has been developed and some successes have been reported[16]. However, this method still applies the structural information in a one-dimensional, 'sequence-like' fashion and fails to take into account the powerful three-dimensional information displayed by protein structures.

In addition, proteins can gain and lose function during evolution and may, indeed, have multiple functions in the cell (Box 1). Sequence-to-function methods cannot specifically identify these complexities. Inaccurate use of sequence-to-function methods has led to significant function-annotation errors in the sequence databases[17].

An alternative approach

An alternative, complementary approach to protein-function prediction uses the sequence-to-structure-to-function paradigm. Here, the goal is to determine the structure of the protein of interest and then to identify the functionally important residues in that structure. Using the chemical structure itself to identify functional sites is more in line with how the protein actually works.

0167-7799/00/$ – see front matter PII: S0167-7799(99)01398-0

In a sense, this is one long-term goal of 'structural genomics' projects[14,15], which are designed to determine all possible protein folds experimentally, just as genome-sequencing projects are determining all protein sequences[16]. This is in contrast to traditional structural-biology approaches, in which one knows the protein's function first and only then, if the function is sufficiently important, determines its structure.

It is implicitly assumed that having the protein's structure will provide insights into its function, thereby furthering the goals of the human-genome-sequencing project. However, knowing a protein's three-dimensional structure is insufficient to determine its function (Box 2). What we really need to analyse and predict the multifunctional aspects of proteins is a method specifically to recognize active sites and binding regions in these protein structures.

Active-site identification

In order to use a structure-based approach to function prediction, one must identify the key residues responsible for a given biochemical activity. For many years, it has been suggested that the active sites in proteins are better conserved than the overall fold. Taken to the limit, this suggests that one could not only identify distant ancestors with the same global fold and the same activity but also proteins with similar functions but distantly related, or possibly unrelated, global folds.

The validity of this suggestion was demonstrated empirically by Nussinov and co-workers, who showed that the active sites of eukaryotic serine proteases, subtilisins and sulfhydryl proteases exhibit similar structural motifs[21]. Furthermore, in a recent modeling study of *Saccharomyces cerevisiae* proteins, protein functional sites were found to be more conserved than other parts of the protein models[22]. Similarly, it has been demonstrated that the catalytic triad of the α/β hydrolases is structurally better conserved than other histidine-containing triads[23]. A comparison of the structure of the hydrolase catalytic triad to other histidine-containing triads shows a distinct bimodal distribution, while a similar analysis done with a randomly selected triad shows a unimodal distribution (Fig. 1).

Kasuya and Thornton[24] generalized this example by creating structural analogs of a few Prosite sequence motifs[16]. For the 20 most-frequently occurring Prosite patterns, the associated local structure is quite distinct. These results provide clear evidence that enzyme active sites are indeed more highly conserved than other parts of the protein.

Identifying active sites in experimental structures

Historically, several groups have attempted to identify functional sites in proteins; these efforts were directed at protein engineering or building functional sites in places where they did not previously exist. This has been successfully accomplished for several metal-binding sites[25–30]. However, highly accurate functional-site descriptors of the backbone and side-chain atoms were required, fueling the belief that significant atomic detail is required in site descriptors for function identification.

Highly detailed residue side-chain descriptors of the active sites of serine proteases and related proteins have been used to identify functional sites[3]. The use of these highly detailed motifs has led to the identification of several novel functional sites in known, high-quality protein structures[3,31]. More automated methods for finding spatial motifs in protein structures have also been described[31,32–40].

Unfortunately, most of these methods require the exact placement of atoms within protein backbones and side chains, and so have not been shown to be reliable to inexact predicted structures. Recently, however, we described the production of fuzzy, inexact descriptors of protein functional sites[13]. As we wish to apply the descriptors to experimental structures as well as to predicted protein models, we used only carbon atoms and side-chain centers-of-mass positions. We call these descriptors 'fuzzy functional forms' (FFFs) and have created them for both the disulfide-oxidoreductase[13,41] and α/β-hydrolase catalytic active sites[13].

The disulfide-oxidoreductase FFF was applied to screen high-resolution structures from the Brookhaven protein database[42]. In a dataset of 364 protein structures, the FFF accurately identified all proteins known to exhibit the disulfide-oxidoreductase active site[13]. In a larger dataset of 1501 proteins, the FFF again accurately identified all proteins with the active site. In addition, it identified another protein, 1fjm, a serine–threonine phosphatase. This result was initially discouraging but subsequent sequence alignment and clustering analysis strongly suggested that this putative site might indeed be a site of redox regulation in the serine–threonine phosphatase-1 subfamily[43]. If confirmed by experiment, this result will highlight the advantages of using structural descriptors to analyse multiple functional sites in proteins. It will also highlight the fact that human

Box 1. Proteins are multifunctional

A common protein characteristic that makes functional analysis based only on homology especially difficult is the tendency of proteins to be multifunctional. For instance, lactate dehydrogenase binds NAD^+, substrate and zinc, and performs a redox reaction. Each of these occurs at different functional sites that are in close proximity and the combination of all four sites creates the fully functional protein.

Other examples of multifunctional proteins are the nucleic-acid-binding proteins. For instance, DNA regulatory proteins often contain a DNA-binding domain, a multimerization domain and additional sites that bind regulatory proteins; a classic example is RecA[99]. The 3C rhinovirus protease exhibits a proteolytic function as well as an RNA-binding function[99,61]. Transcription factors are also complex, multifunctional proteins[62]. It is becoming increasingly important to recognize each of these different functions of gene products of a newly sequenced gene.

The serine–threonine-phosphatase superfamily is a prime example of the difficulties of using standard sequence analysis to recognize the multiple functions found in single proteins. This large protein family is divided into a number of subfamilies, all of which contain an essential phosphatase active site. Subfamilies 1, 2A and 2B exhibit 40% or more sequence identity between them[63]. However, each of these subfamilies is apparently regulated differently in the cell[64–67] and observation suggests that there are different functional sites at which regulation can occur. Because the sequence identity between subfamilies is so high, standard sequence-similarity methods could easily miss/classify new sequences as members of the wrong subfamily if the functional sites are not carefully considered, as was recently demonstrated[43].

These are but a few examples of the multifunctionality of proteins. The recognition of this multifunctional nature is of critical importance to the genomics field. Useful functional-annotation methods must consider all of the specific functions in a given protein and will not just provide a general classification of function.

Box 2. Knowing a protein's structure does not necessarily tell you its function

Because proteins can have similar folds but different functions[68,69], determining the structure of a protein may or may not tell you something about its function. The most well-studied example is the $(\alpha/\beta)_8$ barrel enzymes, of which triose-phosphate isomerase (TIM) is the archetypal representative. Members of this family have similar overall structures but different functions, including different active sites, substrate specificities and cofactor requirements[70,71].

Is this example common? Our own analysis of the 1997 SCOP database[68] shows that the five largest fold families are the ferredoxin-like, the (α/β) barrels, the knottins, the immunoglobulin-like and the flavodoxin-like fold families with 22, 18, 13, 9 and 9 subfamilies, respectively (Fig. I). In fact, 57 of the SCOP fold families consist of multiple superfamilies. These data only show the tip of the iceberg, because each superfamily is further composed of protein families and each individual family can have radically different functions. For example, the ferredoxin-like superfamily contains families identified as Fe–S ferredoxins, ribosomal proteins, DNA-binding proteins and phosphatases, among others.

After this article was submitted, a much-more-detailed analysis of the SCOP database was published[72]. This finds a broad function–structure correlation for some structural classes, but also finds a number of ubiquitous functions and structures that occur across a number of families. The article provides a useful analysis of the confidence with which structure and function can be correlated[72]. Knowing the protein structure by itself is insufficient to annotate a number of functional classes and is also insufficient for annotating the specific details of protein function.

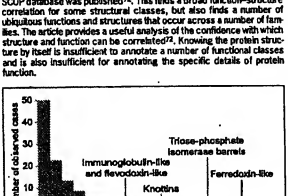


Figure I

Histogram of the numbers of superfamilies found in each SCOP fold family. These data clearly show that proteins with similar structures can have different functions and demonstrate the difficulty of assigning protein function based simply on the three-dimensional structure. The data were taken from the 1997 distribution of SCOP (http://scop.mrc-lmb.cam.ac.uk/scop). For a more-detailed analysis, see Ref. 72.

observation alone is no longer adequate for identifying all functional sites in known protein structures.

To date, the use of structure to identify function has largely focused on high-resolution structures and highly detailed descriptors of protein functional sites. However, the creation of inexact descriptors for functional sites opens the way to the application of these methods to inexact, predicted protein models. The question remains: how good does a model have to be in order to use PPFs to identify its active sites?

The state of the art in structure-prediction methods

For proteins whose sequence identity is above ~30%, one can use homology modeling to build the structure[44]. However, structure prediction is far more difficult for proteins that are not homologous to proteins with known structure. At present, there are two approaches for these sequences: *ab initio* folding[45–48] and threading[49–53].

In *ab initio* folding, one starts from a random conformation and then attempts to assemble the native structure. As this method does not rely on a library of pre-existing folds, it can be used to predict novel folds. The recent CASP3 protein-structure-prediction experiment (http://PredictionCenter.llnl.gov/CASP3) involved the blind prediction of the structure of proteins whose actual structure was about to be experimentally determined. These results indicate that considerable progress has been made[46,54]. For helical and α/β proteins with less than 110 residues, structures were often predicted whose backbone root–mean–square deviation (RMSD) from native ranged from 4–7 Å. Progress is being made with the β proteins, too, although they remain problematic. Because *ab initio* methods can identify novel folds, these methods could be used to help to select sequences likely to yield novel folds in experimental structural-genomics projects.

Another approach to tertiary-structure prediction is threading. Here, for the sequence of interest, one attempts to find the closest matching structure in a library of known folds[52,53]. Threading is applicable to proteins of up to 500 residues or so and is much faster than *ab initio* approaches. However, threading cannot be used to obtain novel folds.

Ab initio *predicted models can be used for automatic protein-function prediction*

The results of the recent CASP3 competition suggest that current modeling methods can often (but not always) create inexact protein models. Are these structures useful for identifying functional sites in proteins? Using the *ab initio* structure-prediction program MONSSTER, the tertiary structure of a glutaredoxin, 1ego, was predicted[46]. For the lowest-energy model, the overall backbone RMSD from the crystal structure was 5.7 Å.

To determine whether this inexact model could be used for function identification, the sets of correctly and incorrectly folded structures were screened with the PFP for disulfide-oxidoreductase activity[55]. The PFP uniquely identified the active site in the correctly folded structure but not in the incorrectly folded ones (Fig. 2). This is a proof-of-principle demonstration that inexact models produced by *ab initio* prediction of structure from sequence can be used for the subsequent prediction of biochemical function. Of course, improvements in the method have to be made before such predictions can be done on a routine basis.

Use of predicted structures from threading in protein-function prediction

At present, practical limitations preclude folding an entire genome of proteins using *ab initio* methods[47]. Threading is more appropriate for achieving the requisite high-throughput structure prediction. Thus, a standard threading algorithm[56] has been used to screen all

proteins in nine genomes for the disulfide-oxidoreductase active site described above.

First, sequences that aligned with the structures of known disulfide oxidoreductases were identified. Then, the structure was searched for matches to the active-site residues and geometry. For those sequences for which other homologs were available, a sequence-conservation profile was constructed[29]. If the putative active-site residues were not conserved in the sequence subfamily to which the protein belongs, that sequence was eliminated. Otherwise, the sequence is predicted to have the function.

Using this sequence-to-structure-to-function method, 99% of the proteins in the nine genomes that have known disulfide-oxidoreductase activity have been found. From 10% to 30% more functional predictions are made than by alternative sequence-based approaches; similar results are seen for the α/β hydrolases[29]. Surprisingly, in spite of the fact that threading algorithms have problems generating good sequence-to-structure alignments, active sites are often accurately aligned, even for very distant matches. This observation would agree with the above experimental results indicating that active sites are well conserved in protein structures.

Importantly, the false-positive rate when using structural information is much lower than that found using sequence-based approaches, as demonstrated by a detailed comparison of the FFF structural approach and the Blocks sequence-motif approach (N. Siew *et al.*, unpublished). In this study, the sequences in eight genomes, including *Bacillus subtilis*, were analyzed for disulfide-oxidoreductase function using the disulfide-oxidoreductase FFF, the thioredoxin Block 00194 and the glutaredoxin Block 00195. If we assume that those sequences identified by both the FFF and Blocks are 'true positives', we find 13 such sequences in the *B. subtilis* genome.

There is no experimental evidence validating all of these 'true positives' and so they are more accurately termed 'consensus positives'. In order to find these 13 'consensus positive' sequences, the FFF hits also had false positives. On the other hand, Blocks hits 23 false positives (Fig. 3). It was previously suggested that the use of a functional requirement adds information to threading and reduces the number of false positives[29]. These data, including the data shown in Fig. 3, validate this claim on a genome-wide basis.

Of course, as no genome has had the function of all of its proteins experimentally annotated, it is impossible to know how many other proteins with the specified biochemical function were not properly identified. This is a critical question for researchers attempting to predict protein function. Experimental confirmation will be needed to validate this or any other method fully. This points out the need for closely coupling computational function-prediction algorithms with experiments.

Weaknesses of using the sequence-to-structure-to-function method of function prediction

Based on studies to date, the identification of enzymatic activity requires a model in which the backbone RMSD from native near the active sites is about 4–5 Å. Predicted models are better at describing the geometry in the core of the molecule than in the loops and so

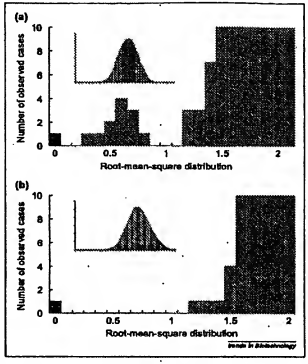


Figure 1

The distribution of root-mean-square distributions (RMSD) between the hydrolase catalytic triad and all other histidine-containing triads shows a bimodal distribution (a); by contrast, the RMSD between a randomly selected (non-catalytic) triad and all other histidine-containing triads has a unimodal distribution (b). The His–Ser–Asp catalytic triad in the protein-1 gpl (Rp2 lipase) (a) and a random histidine-containing triad from 4pga (glutaminase–asparaginase) (b) were structurally aligned to all His-containing triads in a database of 1037 proteins[29]. Actual α/β-hydrolase active sites (a) and the 4pga site (b) are indicated by blue bars; other histidine triads that are not active sites are indicated by red bars. None of the sites found by matching to the 4pga were hydrolase active sites. Inset graphs show the full distribution.

predicting the function of a protein whose active site is in loops may be a problem. Also, the method can currently only be applied to enzyme active sites; substrate- and ligand-binding sites have not been identified using the inexact models. Techniques that will further refine inexact protein models will be quite useful in taking the protein analysis to the next step.

Conclusions

Although sequence-based approaches to protein-function prediction have proved to be very useful, alternatives are needed to assign the biochemical function of the 30–50% of proteins whose function cannot be assigned by any current methods. One emerging approach involves the sequence-to-structure-to-function paradigm. Such structures might be provided by structural-genomics projects or by structure-prediction algorithms. Functional assignment is made by screening the resulting structure against a library of structural descriptors for known active sites or binding regions.

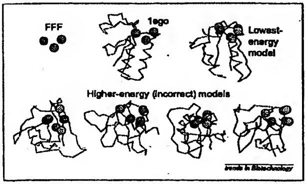


Figure 2

Application of the disulfide-oxidoreductase fuzzy functional form (FFF) to *ab initio* models of glutaredoxin created by the program MONSSTER shows that the FFF can distinguish between correctly folded and misfolded (or higher-energy) models. The FFF is shown as two orange balls (representing the cysteines) and a blue ball (representing the proline). The protein models are shown as magenta wire models with the active-site cysteines and proline shown as yellow and cyan balls, respectively. The FFF clearly distinguishes the correct active site in the crystal structure of the glutaredoxin 1ego and the correctly folded, lowest-energy model. The FFF does not match to the active sites of any of the higher energy, misfolded structures, four of which are shown here.

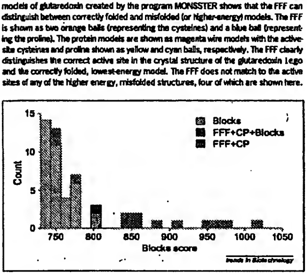


Figure 3

Analysis of the *Bacillus subtilis* genome using the thioredoxin Block 00194. The Blocks score (computed using the publicly available BLIMPS program) is plotted on the x axis and the number of sequences found in each scoring bin is plotted on the y axis. Those sequences identified as 'consensus positives' (identified by both the fuzzy functional form (FFF) and the Block) are shown as red bars. One additional sequence found by the FFF, which is likely to be a true positive, is shown as a blue bar. All other sequences, putative 'false positives', are shown as yellow bars. Using the Blocks score at which all 13 of the 'consensus positives' are found, 23 false positives are also found. In its analysis of the *B. subtilis* genome, the FFF identifies only seven false positives along with the same 13 'consensus positives' (data not shown).

Detailed descriptors will only work on the experimentally determined, high-quality structures. Ideally, however, the descriptors should work on both experimental structures and the cruder models provided by tertiary-structure-prediction algorithms.

The advantages of such an approach are that one need not establish an evolutionary relationship in order to assign function, that more than one function can be assigned to a given protein [an issue of major importance, because proteins are multifunctional (Box 1)] and, ultimately, that having a structure can provide deeper insight into the biological mechanism of protein function and regulation. The disadvantages are that one needs to have the protein's structure before a function can be assigned and that the approach is limited to those functions associated with proteins with at least one solved structure, so that a functional-site descriptor can be constructed.

In this sense, structure-to-function assignment can be thought of as 'functional threading' – find the active-site match in a library of descriptors for known protein active sites. This is the first step in the long process of using structure to assign all levels of function, a goal that is made increasingly important with the emergence of structural genomics. Based on the progress to date, it is apparent that structure will play an important role in the post-genomic era of biology.

Acknowledgment

We thank L. Zhang for producing the data in Box 2 and Fig. 1.

References

1 Oord, F.R.N. and Rodgers, T.M. (1979) Motions in proteins. *Adv. Protein Chem.* 33, 73–165
2 Laskowski, R.A. *et al.* (1996) X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.* 259, 175–201
3 Wallace, A.C. *et al.* (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser–His–Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 5, 1001–1013
4 Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19, 6565–6572
5 Riley, M. (1993) Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952
6 Karp, P.D. and Riley, M. (1993) Representations of metabolic knowledge. *Ismb* 1, 207–215
7 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
8 Pearson, W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.* 266, 227–258
9 Sturrock, S.S. and Collins, J.F. (1993) *Biocomputing Research Unit*, University of Edinburgh, Edinburgh, UK
10 Bairoch, A. *et al.* (1995) The PROSITE database, its status in 1995. *Nucleic Acids Res.* 24, 189–196
11 Henikoff, S. and Henikoff, J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics* 19, 97–107
12 Attwood, T.K. *et al.* (1994) PRINTS – A database of protein motif fingerprints. *Nucleic Acids Res.* 22, 3590–3596
13 Attwood, T.K. *et al.* (1997) Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res.* 25, 212–216
14 Nevill-Manning, C.G. *et al.* (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5865–5871
15 Fetrow, J.S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949–968
16 Yu, L. *et al.* (1998) A homology identification method that combines protein sequence and structure information. *Protein Sci.* 7, 2499–2510
17 Bork, P. and Bairoch, A. (1996) Go hunting in sequence databases but watch out for traps. *Trends Genet.* 12, 425–427
18 Gaasterland, T. (1998) Structural genomics: bioinformatics in the driver's seat. *Nat. Biotechnol.* 16, 625–627
19 McKusick, V.A. (1997) Genomics: structural and functional studies of genomes. *Genomics* 45, 244–249
20 Montelione, G.T. and Anderson, S. (1999) Structural genomics: keystone for a human proteome project. *Nat. Struct. Biol.* 6, 11–12

21 Fischer, D. *et al.* (1994) Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* 3, 769–778

22 Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13597–13602

23 Zhang, L. *et al.* (1998) Functional analysis of *E. coli* proteins for members of the α/β hydrolase family. *Fold. Design* 3, 535–548

24 Kasuya, A. and Thornton, J.M. (1999) Three-dimensional structure analysis of Prosite patterns. *J. Mol. Biol.* 286, 1673–1691

25 Coldren, C.D. *et al.* (1997) The rational design and construction of a cuboidal iron–sulfur protein. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6635–6640

26 Pinto, A.L. *et al.* (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5562–5567

27 Hellinga, H.W. and Richards, F.M. (1991) Construction of new ligand binding sites in proteins of known structure: (I) computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* 222, 763–785

28 Hellinga, H.W. *et al.* (1991) Construction of new ligand binding sites in proteins of known structure: (II) grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* 222, 787–803

29 Klemba, M. and Regan, L. (1995) Characterization of metal binding by a designed protein: single ligand substitutions at a tetrahedral Cys_2His_2 site. *Biochemistry* 34, 10094–10100

30 Klemba, M. *et al.* (1995) Novel metal-binding proteins by design. *Nat. Struct. Biol.* 2, 368–373

31 Farinas, E. and Regan, L. (1998) The *de novo* design of a rubredoxin-like Fe site. *Protein Sci.* 7, 1939–1946

32 Crowder, M.W. *et al.* (1995) Spectroscopic studies on the designed metal-binding sites of the 43C9 single chain antibody. *J. Am. Chem. Soc.* 117, 5627–5634

33 Halfon, S. and Craik, C.S. (1996) Regulation of proteolytic activity by engineered tridentate metal binding loops. *J. Am. Chem. Soc.* 118, 1227–1228

34 Wallace, A.C. *et al.* (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. *Protein Sci.* 6, 2308–2323

35 Kleywegt, G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285, 1887–1897

36 Matsuo, Y. and Nishikawa, K. (1994) Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci.* 3, 2055–2063

37 Russell, R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* 279, 1211–1227

38 Han, K.F. *et al.* (1997) Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* 6, 1587–1590

39 Artymiuk, P.J. *et al.* (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* 236, 327–344

40 Karlin, S. and Zhu, Z.Y. (1996) Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.* 93, 8344–8349

41 Fetrow, J.S. *et al.* (1998) Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703–711

42 Abola, E.E. *et al.* (1987) *Protein Data Bank in Crystallographic Databases: Information Content, Software Systems, Scientific Applications* (Allen, F.H. *et al.*, eds), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester

43 Fetrow, J.S. *et al.* (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* 13, 1866–1874

44 Sali, A. *et al.* (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins* 23, 318–326

45 Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281, 565–577

46 Shortle, D. (1999) The state of the art. *Curr. Biol.* 9, R205–R209

47 Lee, J. *et al.* (1999) Calculation of protein conformation by global optimization of a potential energy function. *Proteins* 3 (Suppl.), 204–208

48 Ortiz, A. *et al.* (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins* 3 (Suppl.), 177–185

49 Bowie, J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170

50 Finkelstein, A.V. and Reva, B.A. (1991) A search for the most stable folds of protein chains. *Nature* 351, 497–499

51 Bryant, S.H. and Lawrence, C.E. (1993) An empirical energy function for threading protein sequence through folding motif. *Proteins* 16, 92–112

52 Lathrop, R. and Smith, T.F. (1996) Global optimum protein threading with gapped alignment and empirical pair scoring function. *J. Mol. Biol.* 255, 641–665

53 Jones, D.T. *et al* (1992). A new approach to protein fold recognition. *Nature* 358, 86–89

54 Sternberg, M.J. *et al.* (1999) Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* 9, 368–373

55 Miller, R.T. *et al.* (1996) Protein fold recognition by sequence threading tools and assessment techniques. *FASEB J.* 10, 171–178

56 Ortiz, A.R. *et al.* (1998) Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* 277, 419–448

57 Skolnick, J. *et al.* (1998) Reduced protein models and their application to the protein folding problem. *J. Biomol. Struct. Dyn.* 16, 381–396

58 Jaroszewski, L. *et al.* (1998) Fold prediction by a hierarchy of sequence, threading and modeling methods. *Protein Sci.* 7, 1431–1440

59 Takahashi, M. *et al.* (1996) Locations of functional domains in the RecA protein: overlap of domains and regulation of activities. *Eur. J. Biochem.* 242, 20–28

60 Leong, L.E. *et al.* (1993) Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5′ noncoding region of the viral RNA: evidence that 3Cpro has different domains for the RNA binding and proteolytic activities. *J. Biol. Chem.* 268, 25735–25739

61 Matthews, D.A. *et al.* (1994) Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site and means for cleaving precursor polyprotein. *Cell* 77, 761–771

62 Ladomery, M. (1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *BioEssays* 19, 903–909

63 Goldberg, J. *et al.* (1995) Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* 376, 745–753

64 Mumby, M.C. and Walter, G. (1993) Protein serine/threonine phosphatases: structure, regulation and functions in cell growth. *Physiol. Rev.* 73, 673–699

65 Jia, Z. (1997) Protein phosphatases: structures and implications. *Biochem. Cell Biol.* 75, 17–26

66 Holmes, C.F.B. and Boland, M.P. (1993) Inhibitors of protein phosphatase-1 and -2A: two of the major serine/threonine protein phosphatases involved in cellular regulation. *Curr. Opin. Struct. Biol.* 3, 934–943

67 Nemani, R. and Lee, E.Y.C. (1993) Reactivity of sulfhydryl groups of the catalytic subunits of rabbit skeletal muscle protein phosphatases 1 and 2A. *Arch. Biochem. Biophys.* 300, 24–29

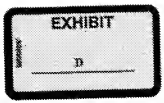
68 Murzin, A.G. *et al.* (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

69 Orengo, C.A. *et al.* (1997) CATH: a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108

70 Lesk, A.M. *et al.* (1989) Structural principles of α/β proteins: the packing of the interior of the sheet. *Proteins Struct. Funct. Genet.* 5, 139–148

71 Farber, G.K. and Petsko, G.A. (1990) The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* 15, 228–234

72 Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288, 147–164



BLAST

Basic Local Alignment Search Tool

•

Edit and Resubmit Save Search Strategies Formatting options Download

Blast 2 sequences

Protein Sequence (806 letters)

Results for: lcl|43901 None(806aa)

Your BLAST job specified more than one input sequence. This box lets you choose which input sequence to show BLAST results for.

Query ID
lcl|43901
lcl|43901
Description
None
Molecule type
amino acid
Query Length
806

Subject ID
43903
Description
None
Molecule type
amino acid
Subject Length
1106
Program
BLASTP 2.2.22+ Citation

Reference
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
Reference - compositional score matrix adjustment
Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.
Other reports: Search Summary [Taxonomy reports] [Multiple alignment] NEW
Search Parameters

Search parameter name Search parameter value

Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F

Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Karlin-Altschul statistics

Params	Ungapped	Gapped
Lambda	0.318991	0.267
K	0.133935	0.041
H	0.404134	0.14

Results Statistics

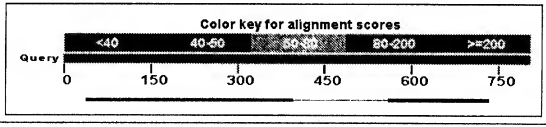
Results Statistics parameter name	Results Statistics parameter value
Effective search space	809244

Graphic Summary

Distribution of 5 Blast Hits on the Query Sequence

[?]

An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple alignments on the same database sequence are connected by a striped line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments. New: This graphic is an overview of database sequences aligned to the query sequence. Alignments are color-coded by score, within one of five score ranges. Multiple alignments on the same database sequence are connected by a dashed line. Mousing over an alignment shows the alignment definition and score in the box at the top. Clicking an alignment displays the alignment detail.



Dot Matrix View

Plot of lcl|43901 vs 43903 [?]

This dot matrix view shows regions of similarity based upon the BLAST results. The query sequence is represented on the X-axis and the numbers represent the bases/residues of the query. The subject is represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Plus strand and protein matches are slanted from the bottom left to the upper right corner, minus strand matches are slanted from the upper left to the lower right. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

Descriptions

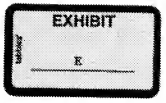
Sequences producing significant alignments:	Score (Bits)	E Value
lcl|43903 unnamed protein product	62.8	1e-13

Alignments Select All Get selected sequences Distance tree of results Multiple alignment NEW

```
>lcl|43903 unnamed protein product
Length=1106

 Score = 62.8 bits (151),  Expect = 1e-13, Method: Compositional matrix adjust.
 Identities = 84/382 (21%), Positives = 153/382 (40%), Gaps = 49/382 (12%)

Query  41   LTILANTTLQITCRGQRDLDWLWPNAQRDSEERVLVTECGGGDSIFCKTLTIPRVVGNDT  100
            L +  ++T  +TC G    + W      +R S+E       D  F    LT+   + G DT
Sbjct  42   LVLNVSSTFVLTCSGSAPVVW-----ERMSQEPPQEM-AKAQDGTFSSVLTLTNLTGLDT  95

Query  101  GAYKCSYRD-----VDIASTVYVYVRDYRSPFIASVSDQHGIVYITENKNKTVVIPCRGS  155
            G Y C++ D     D     +Y++V D    F+ + +++     ++++TE     + IPCR +
Sbjct  96   GEYFCTHNDSRGLETDERKRLYIFVPDPTVGFLPNDAEEL-FIFLTEITE--ITIPCRVT  152

Query  156  ISNLNVSLCARYPEKRFVPDGNRISWDSEIGFTLPSYMISYAGMVFCEAKINDETYQSIM  215
               L V+L  +  +        + +D + GF+       SY     C+  I D    S
Sbjct  153  DPQLVVTLHEKKGDVAL-----PVPYDHQRGFSGIFEDRSY----ICKTTIGDREVDSDA  203

Query  216  YIVVVVGYRIYDVILSPPHEIELSAGEKLVLNCTARTELNVGLDFTWHSPPSKSHHKKIV  275
            Y V  +      +V ++    + +  GE + L C       N  ++F W  P       +K
Sbjct  204  YYVYRLQVSSINVSVNAVQTV-VRQGENITLMCIVIG--NEVVNFEWTYP------RKES  254

Query  276  NRDVKPFPGTVAKM---FLSTLTIESVTKSDQGEYTCVASSGRMIKRNRTFVRVHT--KP  330
              R V+P     +  M    S L I  S     D G YTC  +      ++    + +  KP
Sbjct  255  GRLVEPVTDFLLDMPYHIRSILHIPSAELEDSGTYTCNVTESVNDHQDEKAINITVVESG  314

Query  331  FIAFGSGMKSLVEATVGSQVRIPVKYLSYPAPDIKWYRNGRPIESNYTMIVG--------  382
            ++     + +L  A +     + V  +YP P + W+++ R  +     +
Sbjct  315  YVRLLGEVGTLQFAELHRSRTLQVVFEAYPPPTVLWFKDNRTLGDSSAGEIALSTRNVSE  374

Query  383  -----DELTIMEVTERDAGNYTV  400
                 ELT++ V   +AG+YT+
Sbjct  375  TRYVSELTLVRVKVAEAGHYTM  396


 Score = 40.8 bits (94),  Expect = 4e-07, Method: Compositional matrix adjust.
 Identities = 45/189 (23%), Positives = 75/189 (39%), Gaps = 33/189 (17%)

Query  563  ESVSLLCTADRNTFENLTWYKLGSQATSVHMGESLTPVCKNL-DALWKLNGTMFSNSTND  621
            E+++L+C    N  N  W      ++      G  + PV   L D  + +
Sbjct  229  ENITLMCIVIGNEVVNFEWTYPRKES-----GRLVEPVTDFLLDMPYHIRS---------  274

Query  622  ILIVAFQNASLQDQGDYVCSAQDKKTKKRHCLVKQLILLE----RMAPMITGNLENQTTT  677
               I+    +A L+D G Y C+  +      +  ++E    R+    + G L+
Sbjct  275  --ILHIPSAELEDSGTYTCNVTESVNDHQDEKAINITVVESGYVRLLGEV-GTLQFAELH  331

Query  678  IGETIEVTCPASGNPTPHITWFKDNETLVEDSGIVLRDGNRN---------LTIRRVRKE  728
```

```
              T++V   A   P P + WFKDN TL + S   +     RN          LT+ RV+
Sbjct  332  RSRTLQVVFEAY--PPPTVLWFKDNRTLGDSSAGEIALSTRNVSETRYVSELTLVRVKVA  389

Query  729  DGGLYTCQA  737
            + G YT +A
Sbjct  390  EAGHYTMRA  398


 Score = 20.4 bits (41),  Expect = 0.52, Method: Compositional matrix adjust.
 Identities = 14/67 (20%), Positives = 28/67 (41%), Gaps = 5/67 (7%)

Query  624  IVAFQNASLQDQGDYVCSAQDKK---TKKRHCLVKQLIILERMAPMITGNLENQTTTIGE  680
            ++   N +  D G+Y C+   D +    T +R   L   + + +     +  + E     + E
Sbjct  84   VLTLTNLTGLDTGEYFCTHNDSRGLETDERKRLY--IFVPDPTVGFLPNDAEELFIFLTE  141

Query  681  TIEVTCP  687
              E+T P
Sbjct  142  ITEITIP  148


 Score = 20.4 bits (41),  Expect = 0.59, Method: Compositional matrix adjust.
 Identities = 7/15 (46%), Positives = 8/15 (53%), Gaps = 0/15 (0%)

Query  684  VTCPASGNPTPHITW  698
            V C   G P P+I W
Sbjct  434  VRCRGRGMPQPNIIW  448


 Score = 17.3 bits (33),  Expect = 4.7, Method: Compositional matrix adjust.
 Identities = 8/32 (25%), Positives = 14/32 (43%), Gaps = 0/32 (0%)

Query  162  SLCARYPEKRFVPDGNRISWDSEIGFTLPSYM  193
            +   + +KR  P    S   +G  LPS++
Sbjct  699  TFLQHHSDKRRPPSAELYSNALPVGLPLPSHV  730
```

Select All Get selected sequences Distance tree of results Multiple alignment NEW



BLAST

Basic Local Alignment Search Tool

Edit and Resubmit Save Search Strategies Formatting options Download

Blast 2 sequences

Protein Sequence (806 letters)

Results for: lcl|60337 None(806aa)
Your BLAST job specified more than one input sequence. This box lets you choose which input sequence to show BLAST results for.

Query ID
lcl|60337
lcl|60337
Description
None
Molecule type
amino acid
Query Length
806

Subject ID
60339
Description
None
Molecule type
amino acid
Subject Length
1091
Program
BLASTP 2.2.22+ Citation

Reference
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
Reference - compositional score matrix adjustment
Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.
Other reports: Search Summary [Taxonomy reports] [Multiple alignment] NEW
Search Parameters

Search parameter name	Search parameter value
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F

Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Karlin-Altschul statistics

Params	Ungapped	Gapped
Lambda	0.318991	0.267
K	0.133935	0.041
H	0.404134	0.14

Results Statistics

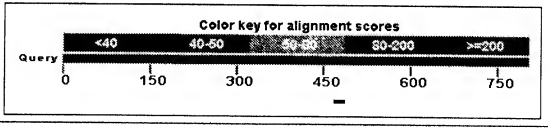
Results Statistics parameter name	Results Statistics parameter value
Effective search space	799624

Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence

[?]

An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple alignments on the same database sequence are connected by a striped line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments. New: This graphic is an overview of database sequences aligned to the query sequence. Alignments are color-coded by score, within one of five score ranges. Multiple alignments on the same database sequence are connected by a dashed line. Mousing over an alignment shows the alignment definition and score in the box at the top. Clicking an alignment displays the alignment detail.



Dot Matrix View

Plot of lcl|60337 vs 60339 [?]

This dot matrix view shows regions of similarity based upon the BLAST results. The query sequence is represented on the X-axis and the numbers represent the bases/residues of the query. The subject is represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Plus strand and protein matches are slanted from the bottom left to the upper right corner, minus strand matches are slanted from the upper left to the lower right. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value
lcl|60339 unnamed protein product	16.5	7.7

Alignments Select All Get selected sequences Distance tree of results Multiple alignment NEW

```
>lcl|60339 unnamed protein product
Length=1091

 Score = 16.5 bits (31),  Expect = 7.7, Method: Compositional matrix adjust.
 Identities = 8/19 (42%), Positives = 8/19 (42%), Gaps = 0/19 (0%)

Query  472   PGQTSPYACKEWRHVEDFQ  490
              P Q  P A  E   V D Q
Sbjct  1071  PSQVLPPASPEGETVADLQ  1089
```

Select All Get selected sequences Distance tree of results Multiple alignment NEW

BLAST

Basic Local Alignment Search Tool

Edit and Resubmit Save Search Strategies Formatting options Download

Blast 2 sequences

Protein Sequence (806 letters)

Results for: lcl|40585 None(806aa)
Your BLAST job specified more than one input sequence. This box lets you choose which input sequence to show BLAST results for.

Query ID
lcl|40585
lcl|40585
Description
None
Molecule type
amino acid
Query Length
806

Subject ID
40587
Description
None
Molecule type
amino acid
Subject Length
820
Program
BLASTP 2.2.22+ Citation

Reference
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
Reference - compositional score matrix adjustment
Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.
Other reports: Search Summary [Taxonomy reports] [Multiple alignment] NEW

Search Parameters

Search parameter name	Search parameter value
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F

Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Karlin-Altschul statistics

Params	Ungapped	Gapped
Lambda	0.318991	0.267
K	0.133935	0.041
H	0.404134	0.14

Results Statistics

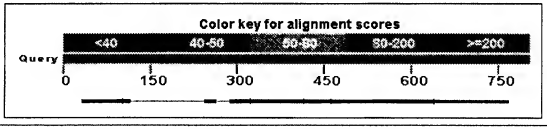
Results Statistics parameter name	Results Statistics parameter value
Effective search space	595935

Graphic Summary

Distribution of 12 Blast Hits on the Query Sequence

[?]

An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple alignments on the same database sequence are connected by a striped line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments. New: This graphic is an overview of database sequences aligned to the query sequence. Alignments are color-coded by score, within one of five score ranges. Multiple alignments on the same database sequence are connected by a dashed line. Mousing over an alignment shows the alignment definition and score in the box at the top. Clicking an alignment displays the alignment detail.



Dot Matrix View :

Plot of lcl|40585 vs 40587 [?]

This dot matrix view shows regions of similarity based upon the BLAST results. The query sequence is represented on the X-axis and the numbers represent the bases/residues of the query. The subject is represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Plus strand and protein matches are slanted from the bottom left to the upper right corner, minus strand matches are slanted from the upper left to the lower right. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value
lcl|40587 unnamed protein product	57.8	3e-12

Alignments Select All Get selected sequences Distance tree of results Multiple alignment NEW

```
>lcl|40587 unnamed protein product
Length=820

 Score = 57.8 bits (138),  Expect = 3e-12, Method: Compositional matrix adjust.
 Identities = 45/165 (27%), Positives = 72/165 (43%), Gaps = 28/165 (16%)

Query  634  DQGDYVCSAQDKKTKKRHCLVKQLILERMA--PMITGNLE-NQTTTIGETIEVTCPASG  690
            D+G+Y C  +++      H    QL ++ER     P++   L  N+T   +G  +E  C
Sbjct  224  DKGNYTCIVENEYGSINHTY--QLDVVERSPHRPILQAGLPANKTVALGSNVEFMCKVYS  281

Query  691  NPTPHITWFKDNET--------------LVEDSGIVLRDGNRN-LTIRRVRKEDGGLYTC  735
            +P PHI W K  E                +++ +G+   D     L +R V  ED G YTC
Sbjct  282  DPQPHIQWLKHIEVNGSKIGPDNLPYVQILKTAGVNTTDKEMEVLHLRNVSFEDAGEYTC  341

Query  736  QACNVLGCARAET-LFIIEGAQEKTN-------LEVIILVGTAVI  772
              A N +G +      L ++E  +E+          LE+II    A +
Sbjct  342  LAGNSIGLSHHSAWLTVLEALEERPAVMTSPLYLEIIIYCTGAFL  386


 Score = 41.6 bits (96),  Expect = 2e-07, Method: Compositional matrix adjust.
 Identities = 34/148 (22%), Positives = 62/148 (41%), Gaps = 8/148 (5%)

Query  325  RVHTKPFIAFGSGMKSLVEATVGSQ-VRIPVKYLSYPADPIKWYRNGRPIESN-----YT  378
            R+   P+       M+  + A    ++ V+          P P ++W +NG+   +      Y
Sbjct  148  RMPVAPYWTSPEKMEKKLHAVPAAKTVKFKCPSSGTPNPTLRWLKNGKEFKPDHRIGGYK  207

Query  379  MIVGDELTIME-VTERDAGNYTVILTNPISMEKQSHMVSLVVNVPPQIGEKALISPMNSY  437
            +        IM+ V   D GNYT I+ N          ++ + +V    P +    +A +     +
Sbjct  208  VRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILQAGLPANKTV  267

Query  438  QYGTMQTLTCTVYANPPLHHIQWYWQLE  465
              G+     C VY++ P  HIQW   +E
Sbjct  268  ALGSNVEFMCKVYSD-PQPHIQWLKHIE  294


 Score = 39.7 bits (91),  Expect = 6e-07, Method: Compositional matrix adjust.
 Identities = 36/152 (23%), Positives = 64/152 (42%), Gaps = 32/152 (21%)

Query  293  TLTIESVTKSDQGEYTCVASSGRMIKRNRTFV-----RVHTKPFIAPGSGMKSLVEATVG  347
            ++ ++SV  SD+G YTC+  +      N T+      R   +P +    +G+ +      +G
Sbjct  214  SIIMDSVVPSDKGNYTCIVEN-EYGSINHTYQLDVVERSPHRPILQ--AGLPANKTVALG  270

Query  348  SQVRIPVKYLSYPAPDIKWYRNGRPIESNYTMIVGDELTIMEVTE----------------  392
            S V    K  S P  I+W ++      IE N + I  D L  +++ +
Sbjct  271  SNVEFMCKVYSDPQPHIQWLKH---IEVNGSKIGPDNLPYVQILKTAGVNTTDKEMEVLH  327

Query  393  ------RDAGNYTVILTNPISMEKQSHMVSLV  418
```

```
                DAG YT +   N I +    S  ++++
Sbjct  328  LRNVSFEDAGEYTCLAGNSIGLSHHSAWLTVL  359
```

Score = 38.9 bits (89), Expect = 1e-06, Method: Compositional matrix adjust.
Identities = 20/67 (29%), Positives = 31/67 (46%), Gaps = 3/67 (4%)

```
Query  679  GETIEVTCPASGNPTPHITWFKDNETLVED---SGIVLRDGNRNLTIRRVRKEDGGLYTC  735
             +T++  CP+SG P P + W K+ +    D    G  +R    ++ +  V    D G YTC
Sbjct  171  AKTVKFKCPSSGTPNPTLRWLKNGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTC  230

Query  736  QACNVLG  742
              N  G
Sbjct  231  IVENEYG  237
```

Score = 33.5 bits (75), Expect = 5e-05, Method: Compositional matrix adjust.
Identities = 59/306 (19%), Positives = 109/306 (35%), Gaps = 53/306 (17%)

```
Query  364  IKWYRNGRPI-ESNYTMIVGDELTIMEVTERDAGNYTVILTNPISMEKQSHMVSLVVNVP  422
            I W R+G   + ESN T I G+E+ + +    D+G Y   + ++P        S      VNV
Sbjct  64   INWLRDGVQLAESNRTRITGEEVEVQDSVPADSGLYACVTSSP----SGSDTTYFSVNVS  119

Query  423  PQIGEKALISPMNSYQYGTMQTLTCTVYANPPLHHIQWYWQLEEACSYR-----------  471
             +           +          +T            P     +   YW     E       +
Sbjct  120  DALPSSEDDDDDDDSSSEEKETDNTKPNRMP----VAPYWTSPEKMEKKLHAVPAAKTVK  175

Query  472  ---PGQTSPYACKEW-RHVEDFQGGNKIEVTKNQYALIEGKNKTVSTLVIQAANVS--AL  525
               P    +P      W ++ ++F+   ++I    K +YA              ++++ +    S
Sbjct  176  FKCPSSGTPNPTLRWLKNGKEFKPDHRIGGYKVRYA--------TWSIIMDSVVPSDKGN  227

Query  526  YKCEAINKAGRGERVISFHVI-RGPEITVQPAAQPTEQ-----ESVSLLCTADRNTFENL  579
            Y C    N+ G              V+ R P    +  A   P  +        +V  +C    +    ++
Sbjct  228  YTCIVENEYGSINHTYQLDVVERSPHRPILQAGLPANKTVALGSNVEFMCKVYSDPQPHI  287

Query  580  TWYKLGSQATSVHM---GESLTPVCKNLDALWKLNGTMFSNSTNDILIVAFQNASLQDQG  636
             W K        H+     G  + P     NL   +  L        + +  ++ ++  +N S +D G
Sbjct  288  QWLK--------HIEVNGSKIGP--DNLPYVQILKTAGVNTTDKEMEVLHLRNVSFEDAG  337

Query  637  DYVCSA  642
            +Y C A
Sbjct  338  EYTCLA  343
```

Score = 24.6 bits (52), Expect = 0.021, Method: Compositional matrix adjust.
Identities = 10/36 (27%), Positives = 20/36 (55%), Gaps = 0/36 (0%)

```
Query  291  LSTLTIESVTKSDQGEYTCVASSGRMIKRNRTFVRV  326
            +  L + +V+  D GEYTC+A +    +   ++ V
Sbjct  323  MEVLHLRNVSFEDAGEYTCLAGNSIGLSHHSAWLTV  358
```

Score = 21.6 bits (44), Expect = 0.21, Method: Compositional matrix adjust.
Identities = 19/75 (25%), Positives = 29/75 (38%), Gaps = 12/75 (16%)

```
Query  49   LQITCRGQRD---LDWLWPNAQRDSEERVLVTECGGGDSIFCKTLTIPRVVGNDTGAYKC  105
            LQ+ CR + D     ++WL       Q       R   +T      G+ +          +     V   D+G Y C
Sbjct  51   LQLRCRLRDDVQSINWLRDGVQLAESNRTRIT----GEEV-----EVQDSVPADSGLYAC  101

Query  106  SYRDVDIASTVYVYV  120
                    + T Y  V
Sbjct  102  VTSSPSGSDTTYFSV  116
```

Score = 21.2 bits (43), Expect = 0.27, Method: Compositional matrix adjust.
Identities = 11/40 (27%), Positives = 16/40 (40%), Gaps = 2/40 (5%)

```
Query  696  ITWFKDNETLVEDSGIVLRDGNRNLTIRRVRKEDGGLYTC  735
            I W +D    L E +      R          + ++      D GLY C
Sbjct  64   INWLRDGVQLAESNRT--RITGEEVEVQDSVPADSGLYAC  101
```

Score = 18.9 bits (37), Expect = 1.4, Method: Compositional matrix adjust.
Identities = 9/22 (40%), Positives = 13/22 (59%), Gaps = 2/22 (9%)

```
Query  247  NCTARTELNVGLDFTWHSPPSK  268
            NCT    EL + +     WH+  PS+
Sbjct  722  NCT--NELYMMMRDCWHAVPSQ  741
```

```
 Score = 18.9 bits (37),  Expect = 1.4, Method: Compositional matrix adjust.
 Identities = 6/16 (37%), Positives = 8/16 (50%), Gaps = 0/16 (0%)

Query  468  CSYRPGQTSPYACKEW  483
            C+ RP  T P   + W
Sbjct  19   CTARPSPTLPEQAQPW  34


 Score = 18.1 bits (35),  Expect = 2.2, Method: Compositional matrix adjust.
 Identities = 18/75 (24%), Positives = 29/75 (38%), Gaps = 12/75 (16%)

Query  37   QKDILTILANTTLQITCRG----QRDLDWLWPNAQRDSEERVLVTECGGGDSIFCKTLTI  92
            +K +  + A  T++  C         L WL    +    + R+       GG  +   T +I
Sbjct  162  EKKLHAVPAAKTVKFKCPSSGTPNPTLRWLKNGKEFKPDHRI------GGYKVRYATWSI  215

Query  93   --PRVVGNDTGAYKC  105
              VV +D G  Y C
Sbjct  216  IMDSVVPSDKGNYTC  230


 Score = 16.5 bits (31),  Expect = 6.2, Method: Compositional matrix adjust.
 Identities = 6/12 (50%), Positives = 8/12 (66%), Gaps = 0/12 (0%)

Query  673  NQTTTIGETIEV  684
            N+T   GE +EV
Sbjct  77   NRTRITGEEVEV  88
```

Select All Get selected sequences Distance tree of results Multiple alignment NEW